



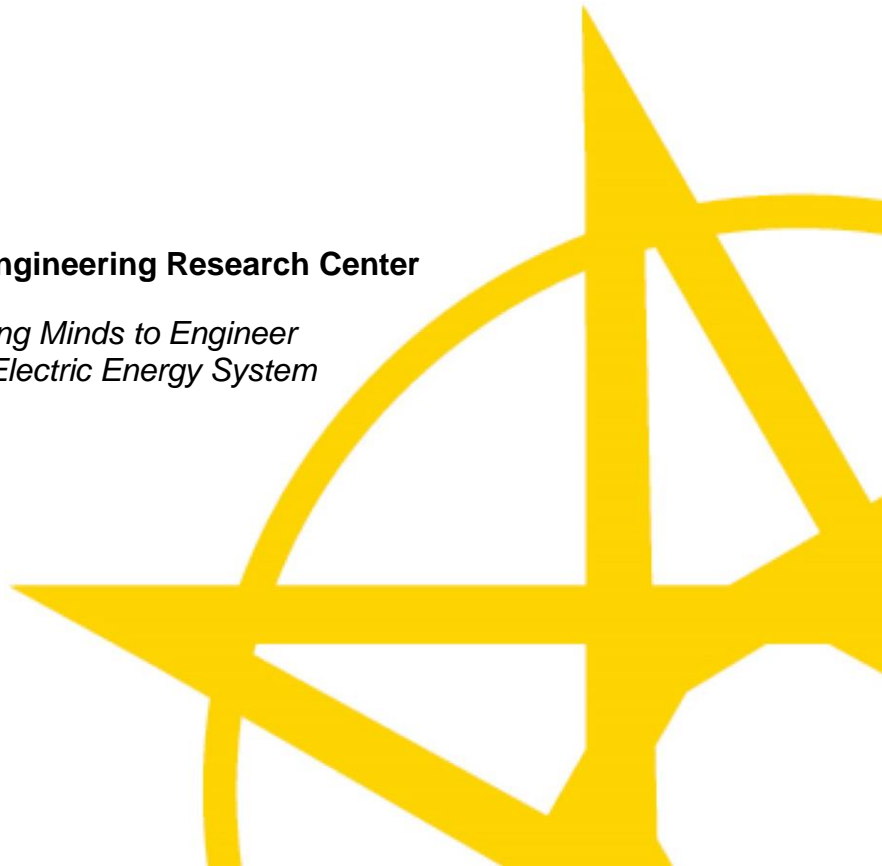
Robust and Decentralized Operations for Managing Renewable Generation and Demand Response in Large-Scale Distribution Systems Simulations

Final Project Report

M-35

Power Systems Engineering Research Center

*Empowering Minds to Engineer
the Future Electric Energy System*



Robust and Decentralized Operations for Managing Renewable Generation and Demand Response in Large-Scale Distribution Systems Simulations

Final Project Report

Project Team

X. Andy Sun, Project Leader
Georgia Institute of Technology

Duncan Callaway
University of California, Berkeley

Graduate Students

Bai Cui
Kaizhao Sun
Georgia Institute of Technology

Felipe Castro
University of California, Berkeley

PSERC Publication 18-12

September 2018

For information about this project, contact:

X. Andy Sun
Georgia Institute of Technology
H. Milton Stewart School of Industrial & Systems Engineering
Atlanta, Georgia 30332-0250
Phone: 404-385-7574
E-mail: andy.sun@isye.gatech.edu

Power Systems Engineering Research Center

The Power Systems Engineering Research Center (PSERC) is a multi-university Center conducting research on challenges facing the electric power industry and educating the next generation of power engineers. More information about PSERC can be found at the Centers website: <http://www.pserc.org>.

For additional information, contact:

Power Systems Engineering Research Center
Arizona State University
577 Engineering Research Center
Tempe, Arizona 85287-5706
Phone: 480-965-1643
Fax: 480-727-2052

Notice Concerning Copyright Material

PSERC members are given permission to copy without fee all or part of this publication for internal use if appropriate attribution is given to this document as the source material. This report is available for downloading from the PSERC website.

©2018 Georgia Institute of Technology. All rights reserved.

Acknowledgments

The authors thank all PSERC members for their technical advice on the project, especially:

Masoud Abbaszadeh, GE Global Research

Jens Boemer, EPRI

Hong Chen, PJM

Bahman Darynian, GE Global Research

Erik Ela, EPRI

Lei Fan, GE Energy

Evangelos Farantatos, EPRI

Eduard Muljadi, NREL

Mirrasoul J. Mousavi, ABB

Jim Price, CAISO

Curtis Roe, ATC

Santosh S. Veda, GE Global Research

Xing Wang, Alstom Grid

Ying Xiao, Alstom Grid

Tongxin Zheng, ISO-NE

Executive Summary

The final report has six chapters. The content of the chapters is summarized below.

1. Chapter 1 develops a robust optimization framework for demand response management. In particular, our model considers the realization uncertainty in demand reduction by a DR resource, which depends on the dispatch decision of the DR resource. This type of intrinsic decision-dependent uncertainty has not been fully recognized and modeled in the DR resource management software. Our work provides a first work in this direction.
2. Chapter 2 proposes a distributed algorithmic framework for solving AC optimal power flow (AC OPF). AC OPF is at the core of power system operations. As the number of distributed generation units and flexible load increases, it is becoming more and more important for the system operators to have the ability to solve AC OPF in a decentralized way. This chapter proposes a rigorous distributed algorithm for solving AC OPF based on the recent progress in convexification of power flow equations. The proposed algorithm shows promising performance in solving real-world sized systems in distributed fashion.
3. Chapter 3 studies the voltage stability issues in the distribution systems. We propose a novel voltage stability-constrained optimal power flow (VSC-OPF) model utilizing a recently proposed sufficient condition on power flow Jacobian nonsingularity. We show that this condition is second-order conic representable when load powers are fixed. Through the incorporation of the convex sufficient condition and thanks to the recent development of convex relaxation of OPF models, we cast a VSC-OPF formulation as a second-order cone program (SOCP). An approximate model is introduced to improve the scalability of the formulation to larger systems. Extensive computation results on MATPOWER and NESTA instances confirm the effectiveness and efficiency of the formulation.
4. Chapter 4 constructs an analytic tool based on a model that captures the interaction between pricing and investment. In contrast to previous approaches, this technique allows consistently comparing portfolios of rates while enabling researchers to model with a significantly greater level of detail the supply side of the sector. A key theoretical implication of the model that underlies this technique is that, by properly updating the portfolio of tariffs, a regulator could induce the welfare maximizing adoption of distributed energy resources and enrollment in rate structures. We develop an algorithm to find globally optimal solutions of this model, which is a nonlinear mathematical program. The results of a computational experiment show that the performance of the algorithm dominates that of commercial non-linear solvers. In addition, to illustrate the practical relevance of the method, we conduct a cost benefit analysis of implementing time-variant tariffs in two electricity systems, California and Denmark. Although portfolios with time-varying rates create value in both systems, these improvements differ enough to advise very different policies. While in Denmark time-varying tariffs appear unattractive, they at least deserve further revision in California. This conclusion is beyond the reach of previous techniques to analyze rates, as they do not capture the interplay between an intermittent supply and a price-responsive demand.

5. In Chapter 5, we develop a technique based on a pricing model that has as a fundamental building block the consumer utility maximization problem. Because researchers do not have to limit themselves to problems with unique solutions, this approach significantly increases the flexibility of the model and, in particular, addresses the limitations of the technique we develop in the first chapter. This gain in flexibility decreases the practicality of our method since the underlying model becomes a Bilevel Problem. To be able to handle realistic instances, we develop a decomposition method based on a non-linear variant of the Alternating Direction Method of Multipliers, which combines Conic and Mixed Integer Programming. A numerical experiment shows that the performance of the solution technique is robust to instance sizes and a wide combination of parameters. We illustrate the relevance of the new method with another applied analysis of rate structures. Our results highlight the value of being able to model in detail distributed energy resources. They also show that ignoring transmission constraints can have meaningful impacts on the analysis of rate structures. In addition, we conduct a distributional analysis, which portrays how our method permits regulators and policy makers to study impacts of a rate update on a heterogeneous population. While a switch in rates could have a positive impact on the aggregate of households, it could benefit some more than others, and even harm some customers. Our technique permits to anticipate these impacts, letting regulators decide among rate structures with considerably more information than what would be available with alternative approaches.
6. In Chapter 6, we conduct an empirical analysis of rate structures in California, which is currently undergoing a rate reform. To contribute to the ongoing regulatory debate about the future of rates, we analyze in depth a set of plausible tariff alternatives. In our analysis, we focus on a scenario in which advanced metering infrastructure and home energy management systems are widely adopted. Our modeling approach allows us to capture a wide variety of temporal and spatial demand substitution patterns without the need of estimating a large number of parameters. We calibrate the model using data of appliance ownership, census household counts, weather patterns, and a model of California's electricity network. The analysis shows that the average gains of implementing time-varying rates with respect to a simple flat rate program are rather mild, not greater than 2 dollars per month, even in the scenario in which volumetric charges are allowed to vary freely from hour to hour. Our results also show that factors such as the presence of an air conditioning system and the exterior temperature profile can have a meaningful impact on the surplus gains that different rates generate on households. These two results combined suggest that defaulting all residential customers into a time-of-use rate structure, which is the current path California is following for the residential sector, may not be an optimal strategy. Targeting different rates to households with different appliance stocks and in different locations will likely be a superior policy.

Project Publications:

- [1] B. Cui and A. Sun. "A New Voltage Stability-Constrained Optimal Power Flow Model: Sufficient Condition, SOCP Representation, and Relaxation," accepted for publication in *IEEE Transactions on Power Systems*, 2018.

- [2] H. Chen, A. Sun, and S. Deng. “Robust Demand Response Portfolio Management under Decision-Dependent Uncertainty,” to be submitted, *INFORMS Journal on Optimization*, 2018.
- [3] K. Sun and A. Sun. “Distributed Algorithms for Solving AC OPF Using SOCP Relaxation,” to be submitted, *IEEE Transactions on Power Systems*, 2018.
- [4] F. Castro and D. Callaway, “Optimal electricity tariff design with demand-side investments,” in preparation.

Student Theses:

- [1] F. Castro. *On Rate Design in Modern Electricity Sectors*, PhD Dissertation, University of California, Berkeley.

Table of Contents

1.	Robust Demand Response Portfolio Management with Decision-Dependent Uncertainty	1
1.1	Introduction	1
1.2	The Deterministic Optimization Model	2
1.2.1	DR Aggregator's Objective Function	2
1.2.2	DR Resource Characterization	3
1.2.3	The Deterministic Optimization Model	4
1.3	The Robust Optimization Model	4
1.3.1	Modeling Operational Uncertainty of DR Resources with Conservativeness Control	4
1.3.2	Robust Demand Response Model	5
1.3.3	Algorithm Framework	9
1.4	Computational Experiments	11
1.4.1	Experiment Setup	11
1.4.2	Computational Analysis	12
1.4.3	Profit Analysis	13
1.4.4	Solution Analysis	14
1.5	Conclusions	16
2.	Decentralized Algorithms for Solving AC-OPF Using SOCP Relaxation	17
2.1	Background	17
2.2	Research Approaches	17
2.2.1	SOCP Relaxation of AC-OPF	18
2.2.2	ADMM	19
2.2.3	Design of Decomposition Scheme	20
2.2.4	Decentralized Algorithm	22
2.3	Numerical Experiments	27
2.3.1	Effect of Different Partitions on ADMM Convergence	27
3.	A New Voltage Stability-Constrained Optimal Power Flow Model: Sufficient Condition, SOCP Representation, and Relaxation	28
3.1	Introduction	28
3.2	Background	29
3.2.1	Notations	29

3.2.2	Power System Modeling.....	29
3.2.3	AC-OPF Formulation	32
3.3	A Sufficient Condition for Nonsingularity of Power flow Jacobian.....	32
3.4	A New Model for VSC-OPF.....	33
3.4.1	New Formulation	33
3.4.2	SOCP Relaxation of VSC-OPF	36
3.4.3	Sparse Approximation of SOCP Relaxation.....	37
3.5	Computational Experiments.....	40
3.5.1	Method.....	41
3.5.2	Results and Discussions.....	42
3.5.3	Comparison with Alternative VSC-OPF Formulation	45
3.6	Conclusions	46
4.	Optimal Rate Design in Modern Electricity Sectors.....	48
4.1	Introduction	48
4.2	Peak-Load Pricing: An Overview	50
4.3	A Method to Compare Rate Structures	51
4.3.1	Consistently Comparing Rate Structures.....	52
4.3.2	A Flexible Cost Function.....	52
4.3.3	Comparing Portfolios of Rate Structures.....	53
4.3.4	Optimal Demand Mix	54
4.3.5	Enhancing the Applicability of the Framework.....	56
4.3.6	Optimal Long Term Incentives	56
4.4	Solution Method.....	57
4.5	An Application: The Value of Real-Time Pricing	59
4.5.1	Analysis Design and Data Assumptions.....	59
4.5.2	Results	61
4.6	Conclusions	62
5.	A Mathematical Programming Approach to Utility Pricing	64
5.1	Introduction	64
5.2	Pricing Utility Services	66
5.2.1	Quantitative Methods for Evaluating Rate Structures	66
5.2.2	Peak-Load Pricing: A Theoretical Framework to Compare Rate Structures	68
5.3	An Alternative Quantitative Technique	69

5.3.1	Limitations of the Model	70
5.3.2	A Realistic Demand Model	71
5.3.3	Examples	72
5.4	Solution Method.....	74
5.4.1	Formulating the Ramsey Problem as an MPEC	74
5.4.2	Decomposing the Problem.....	75
5.4.3	Implementing ADMM.....	77
5.4.4	Testing the Performance of the Approach	84
5.5	A Simple Application.....	86
5.5.1	Designing the Analysis	88
5.5.2	Results	90
5.6	Conclusions	95
6.	Utility Pricing in the Prosumer Era: An Analysis of Residential Electricity Pricing in California	97
6.1	Introduction	97
6.2	California Electricity Sector and the Emergence of Prosumers	98
6.2.1	An Overview of the Sector	98
6.2.2	Residential Rates in California	99
6.2.3	The Emergence of Prosumers	99
6.3	A Modeling Framework to Compare Rate Structures.....	100
6.3.1	Utility Pricing: An Overview of the Theory.....	100
6.3.2	The Regulator's Problem.....	101
6.3.3	A Competitive Wholesale Electricity Market.....	102
6.3.4	The Household Behavior	103
6.3.5	Illustrative Examples	103
6.3.6	Household Heterogeneity and DER Adoption	104
6.3.7	Comparing Rate Structures.....	105
6.4	Modeling California's Electricity Sector	105
6.4.1	Generating Technologies	106
6.4.2	Developing a Model of California's Residential Demand.....	107
6.5	An Analysis of Residential Rate Structures in California.....	108
6.5.1	Aggregated Efficiency Gains.....	110
6.5.2	Implications for Different Households	111

6.5.3	On Carbon Emissions	113
6.6	Conclusions	114

List of Figures

Figure 1.1	The Dynamics of a DR Resource and Realization Uncertainty	3
Figure 1.2	Reduction Plot	14
Figure 1.3	Allocation Plot.....	15
Figure 1.4	Allocation Plot.....	16
Figure 2.1	Accuracy of SOCP Relaxation for IEEE Test Instances	19
Figure 2.2	Comparison of SOCP and SDP Relaxations for IEEE Test Instances up to 3357-Bus System	19
Figure 2.3	Illustration on a 2-Node Graph.....	21
Figure 2.4	Variable Duplication and Constraint Splitting	22
Figure 2.5	Decoupling of Variables	23
Figure 2.6	Numerical Result of Proposed Algorithm	27
Figure 3.1	MSVs of Full and Reduced Power Flow Jacobian with Respect to System Loading for 9-Bus System.....	31
Figure 3.2	Sparsity Pattern of Matrix A for IEEE 300-Bus System	38
Figure 3.3	Results Summary for NESTA Instances From Congested Operating Conditions.....	40
Figure 3.4	Sparse Approximation of NESTA 2737-Bus Test System.....	44
Figure 4.1	Structure of Analysis	59
Figure 4.2	Average Welfare Differences Between Portfolios ((ii) - (i)) by RPS Target. Absolute Differences in Black. Relative Differences in Gray. Line Thickness Represents Different Elasticity Levels. Thicker Lines Correspond to Lower Own- and Cross-Price Elasticities	61
Figure 4.3	Net-Load Duration Curves by RPS Target. California Net-Load in Black. Denmark Net-Load in Gray. All Curves Reach a Maximum of 100%	63
Figure 5.1	Primal and Dual Residual Evolution Across Iterations	86
Figure 5.2	Iteration Metrics Versus Instances Sizes	87
Figure 5.3	Structure of Analysis. An Instance Corresponds to the Combination of a Rate Structure, Network Parameter and the Renewable and DER Costs. There Are 30 Instances in Total.....	88
Figure 5.4	Network Model. The Letter X Denotes the Reactance of the Branch, and the Arrow the Default Direction of the Flows. Buses 1 and 2 Correspond to Loads, and 3, 4 and 5 to Generating Technologies	89
Figure 5.5	Household Configuration per Bus. There Are in Total 6 Different House- holds Types. The Final Number per Type, α , Is an Outcome of the Model, and is Constrained by the Number of Representative Households	90

Figure 5.6	Time Series.	91
Figure 6.1	Distribution of Households Across Net Surplus Gains	112

List of Tables

Table 1.1	Resource Property	11
Table 1.2	Cost Property in a DR Event	11
Table 1.3	Algorithm Performance	12
Table 1.4	ADA vs MIP in Computation Time with Demand Level at 80%(N= 50).....	13
Table 1.5	ADA vs MIP in Computation Time with Demand Level at 30% (N = 50)	13
Table 3.1	Results Summary for Standard IEEE Instances.	40
Table 3.2	Results Summary of Sparse Approximation for Large NESTA Instances From Congested Operating Conditions.	45
Table 3.3	Comparison of the Effect of Voltage Stability Improvement of Different VSC-OPF Formulations.	46
Table 4.1	Economic Parameters of Supply-Side Technologies	60
Table 4.2	Demand Elasticities.....	60
Table 4.3	Demand Mix by RPS Target	62
Table 5.1	Economic Parameters of Supply-Side Technologies	88
Table 5.2	Welfare Gains with Respect to FR in Percentages.....	92
Table 5.3	Rooftop Solar PV Adoption and Generation Fleet Utilization.....	93
Table 5.4	Net Surplus Increase per Household with Respect to FR [%].....	94
Table 6.1	Generating Technologies Costs and GHG Emissions	107
Table 6.2	Benefits and Costs: Changes with Respect to Flat Rate Structure	111
Table 6.3	Capacity, Production and Emissions Changes with Respect to FR Scenario	114

1. Robust Demand Response Portfolio Management with Decision-Dependent Uncertainty

1.1 Introduction

The increasing deployment of smart grid technologies has greatly enhanced the flexibility and responsiveness of electricity demand. As the trend develops, demand response (DR) is gradually becoming an important and valuable resource in today's electricity markets [3, 84]. Many utilities and energy service companies aggregate DR resources into portfolios and participate in electricity markets' demand response programs. Such DR portfolios can have thousands of DR resources of various characteristics [65]. Optimal management of these large-scale DR portfolios is an important task facing DR aggregators.

There is a rapidly growing body of literature on demand response scheduling, see [3, 64, 84, 119, 125, 126, 142, 154] for recent surveys. As a small sampling of the recent literature: Tsui and Chan propose a deterministic optimization model to solve the automatic load management problem in a smart home [141]. Karangelos and Bouffard in [79] develop a forward market clearing algorithm to resolve the demand flexibility problem with the goal to co-optimize the scheduling cost and security of the system. Rastegar and Fotuhi-Firuzabad [121] discussed a novel control approach based on online optimization algorithm to manage the operations of responsive electrical appliances.

DR resources are very different from conventional generation resources. In particular, they can have significant uncertainty in their availability and operational consistency. Also, the market environment has intrinsic uncertainty in electricity prices. To optimally manage the performance of DR resources, these uncertain factors must be considered.

The impact of uncertainty in electricity price has been extensively studied. For example, various robust optimization models that consider price uncertainty are proposed in [34, 37, 57, 82]. In particular, adaptive robust optimization models involving two stages have been discussed by [5, 8, 157]. Stochastic programming models and Markov decision models are studied for price uncertainty and residential demand management in [34, 83]. Optimal DR contract design under price uncertainty is also studied, e.g. see [26]. DR resources scheduling is also studied in the context of unit commitment from a system operator's perspective [112, 148].

The existing literature has extensively studied the DR scheduling problem involving residential customers under price uncertainty. In this paper, we focus on large-scale DR portfolio management of commercial and industrial (C&I) resources and deal with another significant type of uncertainty. In particular, our study makes contributions to the DR portfolio management problem in the following three key aspects.

1. The first aspect is the modeling of characteristic dynamics of a DR event involving C&I resources. In particular, each C&I resource has a specific set of operating requirements, such as demand reduction capacity, ramping rates, schedule smoothness, etc. We propose a deterministic MIP model for the optimal management of a DR portfolio, which properly characterizes these complex operating characteristics of DR resources.

2. The second aspect is to consider *operational uncertainty* of DR resources. More specifically, the realized demand reduction of a DR resource during a DR event can have significant, uncertain deviation from the scheduled action. This type of operational uncertainty can depend on the scheduled action and has a different nature from the exogenous uncertainty such as price uncertainty discussed in the literature. Robust optimization problem with decision-dependent uncertainty set has been a challenging problem, as is characterized by [89] and [106]. Both [89] and [106] investigate the uncertainty sets induced by a binary vector with finite dimension. Different from their discussion, this paper formulates a robust optimization model for DR portfolio management together with an uncertainty set that depends on the continuous reduction decision variables for the DR resources. Despite that the continuous decision variables can induce uncountably many uncertainty sets, we show that exploiting the special structure of the uncertainty set admits a mixed integer programming formulation.
3. The third aspect is on solution methodology. The proposed DR portfolio management models naturally involve discrete decisions. Although robust optimization model can also be formulated as a MIP problem. Off-the-shelf optimization software is not sufficient to deal with the large-scale math programming formulation. We develop a warm-start framework that implements the improved alternating-direction algorithm for the equivalent MINLP problem, which demonstrate numerical computation quality.

The paper is organized as follows. Section 1.2 introduces the deterministic optimization model and discusses specific DR dynamics. Section 1.3 proposes the robust optimization extension and introduces the new uncertainty sets that model operational uncertainty in DR resources. Section 1.4 develops an improved alternating direction algorithm. Section 1.5 reports computational experiments and analysis. Section 1.6 concludes the paper.

1.2 The Deterministic Optimization Model

1.2.1 DR Aggregator's Objective Function

There are three main players in a DR event: the system operator, the DR aggregator, and the DR resources contracted with the DR aggregator. Details of the specific contracts may differ, but the basic feature is that the DR aggregator gains revenue from the system operator for providing the required demand reduction, at the same time, it offers payment to the participating DR resources in its portfolio [65].

We take the perspective of the DR aggregator, and model its objective function as the profit obtained from the market revenue minus potential penalties. In particular, we set c_i to be the income earned from dispatching DR resource i for a unit demand reduction. If the total reduction level from DR resources is less than the required level at time t , the DR aggregator suffers from an under-commitment unit cost s_t caused by refund, contractual penalty and some other invisible cost such as severe loss of market. If the total load reduction level is above the required level, the DR aggregator suffers from the over-commitment unit cost h_t as the system operator suffers from value loss of DR resources. The DR aggregator makes a clear statement that neither over-commitment nor under-commitment benefits the operations of the company, which implies the relation $s_t \geq c_i$ and $h_t \geq c_i$ for all i, t .

To model the above profit structure, we propose a Newsvendor objective. Let D_t be the required total demand reduction level at time t , which is a deterministic parameter known to the DR aggregator. Let $\mathbf{p} = (p_i^t)$ be the vector where p_i^t is the demand reduction level of resource i at the beginning of time t . As a convention, we use positive value of p_i^t for demand reduction. We define $\mathbf{x} = (x_i^t)$ as the vector of x_i^t to denote whether resource i is committed at time t . We let $\mathbf{w} = (w_i^t)$ be such that $w_i^t = 1$ if $p_i^t - p_i^{t-1} \geq 0$, and let $\mathbf{v} = (v_i^t)$ be such that $v_i^t = 1$ if $p_i^t - p_i^{t-1} \leq 0$.

We define the objective $f(\mathbf{p})$ as

$$\sum_t \left[h_t \left(\sum_i p_i^t - D_t \right)^+ + s_t \left(D_t - \sum_i p_i^t \right)^+ - \sum_i c_i p_i^t \right], \quad (1.1)$$

where $(x)^+ := \max(x, 0)$. This objective function can be also understood as a penalization method to avoid infeasibility in exactly matching the required demand reduction levels at all times. Such infeasibility can happen as the DR scheduling problem involves complicated operational constraints, which are shown next.

1.2.2 DR Resource Characterization

Each DR resource has a set of operational characteristics, which have to be respected during a DR event. Figure 1.1 illustrates these key characteristics on a scheduled dispatch trajectory of a DR resource.

The key characteristics are explained below.

1. Reduction constraints: each DR resource has a capacity p_i^{\max} and minimum commitment requirement p_i^{\min}
2. Ramping constraints: DR resources have their ramping limits r_i^+ and r_i^-
3. Smoothness constraints: every time a DR resource reduces its demand, its demand level cannot increase again before at least T_{eup} periods. This is to respect the inertia in the DR resource, similarly for the decreasing smoothness constraint.

These DR characteristics makes the resulting model different from a unit commitment (UC) problem for conventional generation scheduling.

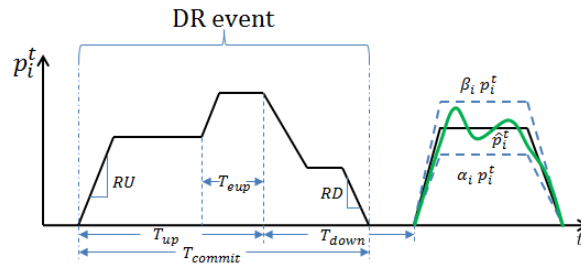


Figure 1.1: The Dynamics of a DR Resource and Realization Uncertainty

1.2.3 The Deterministic Optimization Model

We propose the following deterministic model for DR portfolio management.

$$\begin{aligned}
& \min_{\mathbf{p}, \mathbf{x}, \mathbf{w}, \mathbf{v}} f(\mathbf{p}) & (1.2a) \\
\text{s.t. } & p_i^{\min} x_i^t \leq p_i^t \leq p_i^{\max} x_i^t \quad \forall i, t, & (1.2b) \\
& -r_i^- x_i^t \leq p_i^{t+1} - p_i^t \leq r_i^+ x_i^{t+1} \quad \forall i, t, & (1.2c) \\
& p_i^t - p_i^{t-1} \leq M w_i^t \quad \forall i, t, & (1.2d) \\
& p_i^\tau - p_i^{\tau-1} \geq -M(1 - w_i^t) \quad \forall i, t & \\
& \quad \forall \tau = t, \dots, \min(t + T_{\text{eup}}^i - 1, T), & (1.2e) \\
& p_i^t - p_i^{t-1} \geq -M v_i^t \quad \forall i, t, & (1.2f) \\
& p_i^\tau - p_i^{\tau-1} \leq M(1 - v_i^t) \quad \forall i, t, & \\
& \quad \forall \tau = t, \dots, \min(t + T_{\text{edn}}^i - 1, T) & (1.2g) \\
& x_i^t, w_i^t, v_i^t \in \{0, 1\} \quad \forall i, t. & (1.2h)
\end{aligned}$$

The objective function $f(\mathbf{p})$ is defined in (2.1). In constraint (1.2b), when a DR resource is committed i.e. $x_i^t = 1$, the reduction amount p_i^t is subject to the upper and lower bounds. Constraint (1.2c) defines the maximum up and down ramping rates for committed resource i at time t . As discussed earlier, the DR aggregator needs to respect smoothness characteristics in scheduling demand reduction. In constraints (1.2d) and (1.2e), if resource i increases its dispatch at any time t , it has to keep the non-decreasing trend for a minimum of T_{eup}^i periods. Similarly, constraints (1.2f) and (1.2g) require that if resource i decreases its commitment at any time t , it has to keep the non-increasing trend at least for the next T_{edn}^i periods. In summary, the proposed model (1.2) respects the DR dynamics both during a commitment cycle and between adjacent commitment cycles of a DR resource. The objective function is a piecewise linear convex function and can be easily linearized.

1.3 The Robust Optimization Model

1.3.1 Modeling Operational Uncertainty of DR Resources with Conservativeness Control

In a DR event, the DR aggregator schedules the reduction level for each DR resource. However, unlike conventional generators, DR resources can have significant uncertainty in their demand reduction performance due to unexpected factors in actual operations and market conditions. The final realized reduction level of a DR resource may be quite different from the scheduled level.

We model the final realization of the demand reduction as $\tilde{p}_i^t = p_i^t + \Delta p_i^t$, where Δp_i^t can be viewed as the “implementation error” between the realized demand reduction \tilde{p}_i^t and the scheduled amount p_i^t . The implementation error Δp_i^t can be either negative or positive. To model the uncertainty for

Δp_i^t under budget constraint, we propose an operational uncertainty set \mathcal{U} represented by

$$\mathcal{U}_t(\mathbf{p}^t, \alpha, \beta) = \left\{ \Delta \mathbf{p}^t = (\Delta p_1^t, \dots, \Delta p_N^t) : \alpha_i p_i^t \leq \Delta p_i^t \leq \beta_i p_i^t, \sum_{i=1}^N |\Delta p_i^t| \leq \Gamma^t \sum_{i=1}^N p_i^{\max}, \forall i \right\}. \quad (1.3)$$

Parameters in the definition satisfy $\alpha_i \leq 0$ and $\beta_i \geq 0$ for all i . A key feature of this uncertainty set is that the upper and lower bounds on Δp_i^t are not fixed numbers. Instead, they are functions of the scheduled reduction level p_i^t . This dependence on decision variables distinguishes the proposed uncertainty sets from traditional ones, and offers a way to model the situation where the magnitude of the uncertainty range depends on the demand reduction level. Also, there is no realization uncertainty if the scheduled demand reduction is zero, i.e. $p_i^t = 0$. In this uncertainty set, we use $\Gamma^t \geq 0$ to denote the level of aggregator's ability to control the operations uncertainty. It depends on factors such as budget, willingness to take risk, the market stability, etc.

1.3.2 Robust Demand Response Model

Now we introduce the following robust optimization model, which finds a robust DR schedule that minimizes the realized operational cost in the worst-case scenario under DR operational uncertainties.

$$\min_{\mathbf{p}} \max_{\Delta \mathbf{p} \in \mathcal{U}(\mathbf{p})} f(\mathbf{p}, \Delta \mathbf{p}) \quad (1.4a)$$

$$\text{s.t. } \mathbf{p} \in \Omega, \quad (1.4b)$$

where $f(\mathbf{p}, \Delta \mathbf{p})$ is given as

$$\sum_t \left\{ h_t \left(\sum_i (p_i^t + \Delta p_i^t) - D_t \right)^+ + s_t \left(D_t - \sum_i (p_i^t + \Delta p_i^t) \right)^+ - \sum_i c_i (p_i^t + \Delta p_i^t) \right\}. \quad (1.5)$$

The feasible set Ω is defined in the capacity and ramping constraints of the deterministic model, i.e. $\Omega = \{\mathbf{p} : \exists \mathbf{x}, \mathbf{w}, \mathbf{u} \text{ satisfies (1.2b)-(1.2h)}\}$. We have $\mathcal{U}(\mathbf{p})$ defined in (1.3), where we suppress the parameters α, β, Γ to save space.

The problem has a min-max structure with a nonlinear objective function $f(\mathbf{p}, \Delta \mathbf{p})$ defined in (1.5). In the following result, we demonstrate a more tractable formulation for optimization problem (1.4).

Proposition 1. *In the robust optimization model, $\min_{\mathbf{p} \in \Omega} \max_{\Delta \mathbf{p} \in \mathcal{U}(\mathbf{p})} f(\mathbf{p}, \Delta \mathbf{p})$ can be reformulated by*

$$\min_{\mathbf{Y}, \mathbf{p}} \sum_t Y_t \quad (1.6a)$$

$$Y_t \geq \max_{\Delta \mathbf{q}_1^t \in \mathcal{U}_t(\mathbf{p}_t)} \left[\sum_{i=1}^N (h_t - c_i) \Delta q_{1i}^t \right] + \sum_i (h_t - c_i) p_i^t - h_t D_t, \forall t \quad (1.6b)$$

$$Y_t \geq \max_{\substack{\Delta \mathbf{q}_2^t \in \mathcal{U}_t(\mathbf{p}_t) \\ \mathbf{p} \in \Omega}} \left[- \sum_{i=1}^N (s_t + c_i) \Delta q_{2i}^t \right] - \sum_i (s_t + c_i) p_i^t + s_t D_t, \quad \forall t \quad (1.6c)$$

$$\mathbf{p} \in \Omega \quad (1.6d)$$

Proof. Similar to the structure in the deterministic model, we implement a news-vendor type of objective where the decision variable p_i^t becomes $\tilde{p}_i^t = p_i^t + \Delta p_i^t$ under the operations uncertainty. The robust optimization model can be written compactly as:

$$Z = \min_{\mathbf{p} \in \Omega} \max_{\Delta \mathbf{p} \in \mathcal{U}(\mathbf{p})} \sum_t \left\{ h_t \left(\sum_i (p_i^t + \Delta p_i^t) - D_t \right)^+ + s_t \left(D_t - \sum_i (p_i^t + \Delta p_i^t) \right)^+ - \sum_i c_i (p_i^t + \Delta p_i^t) \right\}.$$

Step 1: derive an equivalent expression. We first show that the original objective is equivalent to

$$Z = \min_{\mathbf{p} \in \Omega} \sum_{t=1}^T \max_{\Delta \mathbf{p}^t \in \mathcal{U}_t(\mathbf{p}, \Gamma)} \max \left\{ (h_t - c_i)(p_i^t + \Delta p_i^t) - h_t D_t, -(s_t + c_i)(p_i^t + \Delta p_i^t) + s_t D_t \right\}.$$

To prove this claim, we denote $X_t(\Delta \mathbf{p}^t)$, Y and Y' as

$$\begin{aligned} X_t(\Delta \mathbf{p}^t) &= h_t \left(\sum_i (p_i^t + \Delta p_i^t) - D_t \right)^+ + s_t \left(D_t - \sum_i (p_i^t + \Delta p_i^t) \right)^+ - \sum_i c_i (p_i^t + \Delta p_i^t) \\ &= \max \left\{ (h_t - c_i)(p_i^t + \Delta p_i^t) - h_t D_t, -(s_t + c_i)(p_i^t + \Delta p_i^t) + s_t D_t \right\}, \\ Y &= \max_{\Delta \mathbf{p} \in \mathcal{U}} \sum_t X_t(\Delta \mathbf{p}^t), \\ Y' &= \sum_t \max_{\Delta \mathbf{p}^t \in \mathcal{U}_t} X_t(\Delta \mathbf{p}^t). \end{aligned}$$

Using this notation, it is sufficient to prove that $Y = Y'$. The expression of Y and Y' suggests that $Y \leq Y'$. Suppose towards contradiction that $Y < Y'$, then there exists $\Delta \mathbf{p}^\tau \in \mathcal{U}_\tau$ for all $\tau = 1, \dots, T$ such that $Y < \sum_t X_t(\Delta \mathbf{p}^t)$. However, vector $\Delta \mathbf{p}' = (\Delta \mathbf{p}^1, \Delta \mathbf{p}^2, \dots, \Delta \mathbf{p}^T)$ satisfies that $\Delta \mathbf{p}' \in \mathcal{U}$ and $Y \geq \sum_t X_t(\Delta \mathbf{p}^t)$, which is a contradiction to the assumption. Thus, we have $Y = Y'$.

Step 2: conclude the claim In this step, we want to show that the reformulation is equivalent to

$$Z = \min_{\mathbf{p} \in \Omega} \sum_{t=1}^T \max \left\{ \max_{\Delta \mathbf{q}_1^t \in \mathcal{U}_t(\mathbf{p})} f_t(\mathbf{p}, \Delta \mathbf{q}_1^t), \max_{\Delta \mathbf{q}_2^t \in \mathcal{U}_t(\mathbf{p})} g_t(\mathbf{p}, \Delta \mathbf{q}_2^t) \right\},$$

where $f_t(\mathbf{p}, \Delta \mathbf{q}_1^t) = \sum_i (h_t - c_i) \Delta q_{1i}^t + \sum_i (h_t - c_i) p_i^t - h_t D_t$ and $g_t(\mathbf{p}, \Delta \mathbf{q}_2^t) = - \sum_i (s_t + c_i) \Delta q_{2i}^t - \sum_i (s_t + c_i) p_i^t + s_t D_t$. We denote

$$W_t = \max_{\Delta \mathbf{p}^t \in \mathcal{U}_t} \max \left\{ f_t(\mathbf{p}, \Delta \mathbf{p}^t), g_t(\mathbf{p}, \Delta \mathbf{p}^t) \right\},$$

$$W'_t = \max \left\{ \max_{\Delta \mathbf{q}_1^t \in \mathcal{U}_t(\mathbf{p})} f_t(\mathbf{p}, \Delta \mathbf{q}_1^t), \max_{\Delta \mathbf{q}_2^t \in \mathcal{U}_t(\mathbf{p})} g_t(\mathbf{p}, \Delta \mathbf{q}_2^t) \right\}.$$

It is sufficient to show that $W_t = W'_t$. The expression directly suggests that $W_t \leq W'_t$. Suppose towards contradiction that $W_t < W'_t$, then there exists $\Delta \mathbf{q}^t \in \mathcal{U}_t(\mathbf{p})$ such that $\max \left\{ f_t(\mathbf{p}, \Delta \mathbf{p}^t), g_t(\mathbf{p}, \Delta \mathbf{p}^t) \right\} < \max \left\{ f_t(\mathbf{p}, \Delta \mathbf{q}^t), g_t(\mathbf{p}, \Delta \mathbf{q}^t) \right\}$ for all $\Delta \mathbf{p}^t \in \mathcal{U}_t(\mathbf{p})$. However, by setting $\Delta \mathbf{q}^t$ is only a feasible solution in W_t , which suggests that $\max \left\{ f_t(\mathbf{p}, \Delta \mathbf{q}^t), g_t(\mathbf{p}, \Delta \mathbf{q}^t) \right\} \leq \max \left\{ f_t(\mathbf{p}, \Delta \mathbf{p}^t), g_t(\mathbf{p}, \Delta \mathbf{p}^t) \right\}$. The contradiction suggests that $W_t = W'_t$.

To conclude the proof, we use an equivalent representation for the optimization problem

$$\begin{aligned} Z &= \min_{\mathbf{Y}, \mathbf{p}, \Delta \mathbf{q}} \sum_t Y_t \\ Y_t &\geq \max_{\Delta \mathbf{q}_1^t \in \mathcal{U}_t} \left[\sum_i (h_t - c_i) \Delta q_{1i}^t \right] + \sum_i (h_t - c_i) p_i^t - h_t D_t \\ Y_t &\geq \max_{\Delta \mathbf{q}_2^t \in \mathcal{U}_t} \left[- \sum_i (s_t + c_i) \Delta q_{2i}^t \right] - \sum_i (s_t + c_i) p_i^t + s_t D_t \\ \mathbf{p} &\in \Omega \end{aligned}$$

□

Proposition 1 implies that our robust optimization model can be treated as a two-stage problem where the second stage includes a total of $2T$ linear programs defined in (1.6b) and (1.6c). For example, a second-stage problem $\max_{\Delta \mathbf{q}_1^t \in \mathcal{U}_t(\mathbf{p}_t)} [\sum_{i=1}^N (h_t - c_i) \Delta q_{1i}^t]$ for period $t \in \{1, \dots, T\}$ in (1.6b) has the following linear formulation given vector \mathbf{p}

$$\max_{\Delta \mathbf{q}_1^t, \mathbf{q}_1^t} \sum_{i=1}^N (h_t - c_i) \Delta q_{1i}^t \quad (1.7a)$$

$$\text{s.t.} \quad \alpha_i p_i^t \leq \Delta q_{1i}^t \leq \beta_i p_i^t, \quad \forall i \quad (1.7b)$$

$$-q_{1i}^t \leq \Delta q_{1i}^t \leq q_{1i}^t, \quad \forall i \quad (1.7c)$$

$$\sum_{i=1}^N q_{1i}^t \leq \Gamma^t \sum_{i=1}^N p_i^{\max} \quad (1.7d)$$

One challenge for this problem is that the uncertainty set is adaptive to decision variables, which makes the traditional techniques not applicable. To resolve the challenge, we redefine some parameters in the second-stage problem in (1.6b). Let $\tilde{c}_j^t = |h_t - c_j|$ and $\gamma_j^t = |\alpha_j| \mathbb{1}_{h_t - c_j \leq 0} + |\beta_j| \mathbb{1}_{h_t - c_j > 0}$. Define a bijective ranking mapping $\phi_t : \{1, \dots, N\} \rightarrow \{1, \dots, N\}$ that satisfies $\tilde{c}_{\phi_t(i)}^t \geq \tilde{c}_{\phi_t(j)}^t$ if and only if $i \leq j$. Similarly, we redefine the parameters in the second-stage problems in (1.6c): $\tilde{d}_j^t = |-(s_t + c_j)|$, $\eta_j^t = |\alpha_j| \mathbb{1}_{-(s_t + c_j) \leq 0} + |\beta_j| \mathbb{1}_{-(s_t + c_j) > 0}$, and the bijective ranking mapping $\psi_t : \{1, \dots, N\} \rightarrow \{1, \dots, N\}$ satisfying $\tilde{d}_{\psi_t(i)}^t \geq \tilde{d}_{\psi_t(j)}^t$ if and only if $i \leq j$. We

also denote $\zeta_t = \Gamma^t \sum_{i=1}^N p_i^{\max}$ for all t .

We show that the optimal values of the second-stage problem in (1.6b) and (1.6c) can be explicitly formulated as the minimum of finite piece-wise linear functions of decision variable \mathbf{p} .

Theorem 1. *The optimal value for the maximization problems in (1.6b) and (1.6c) satisfy*

$$\begin{aligned} & \max_{\Delta \mathbf{q}_1^t \in \mathcal{U}_t(\mathbf{p}_t)} \left[\sum_{i=1}^N (h_t - c_i) \Delta q_{1i}^t \right] \\ &= \min \left\{ \left\{ \sum_{i=1}^N \tilde{c}_i^t \gamma_i^t p_i^t \right\} \cup \left\{ \sum_{i=1}^{k-1} (\tilde{c}_{\phi_t(i)}^t - \tilde{c}_{\phi_t(k)}^t) \gamma_{\phi_t(i)}^t p_{\phi_t(i)}^t + \tilde{c}_{\phi_t(k)}^t \zeta_t, \quad k = 1, \dots, N \right\} \right\}, \\ & \max_{\Delta \mathbf{q}_2^t \in \mathcal{U}_t(\mathbf{p}_t)} \left[\sum_{i=1}^N (-s_t - c_i) \Delta q_{2i}^t \right] \\ &= \min \left\{ \left\{ \sum_{i=1}^N \tilde{d}_i^t \eta_i^t p_i^t \right\} \cup \left\{ \sum_{i=1}^{k-1} (\tilde{d}_{\psi_t(i)}^t - \tilde{d}_{\psi_t(k)}^t) \eta_{\psi_t(i)}^t p_{\psi_t(i)}^t + \tilde{d}_{\psi_t(k)}^t \zeta_t, \quad k = 1, \dots, N \right\} \right\}. \end{aligned}$$

The theorem implies that the robust optimization problem can be formulated as the following MIP with $2T$ sets of disjunctive constraints.

$$\min_{\mathbf{Y}, \mathbf{p}} \sum_t Y_t \tag{1.8a}$$

$$Y_t \geq \sum_{i=1}^N \tilde{c}_i^t \gamma_i^t p_i^t + \sum_i (h_t - c_i) p_i^t - h_t D_t - M \mu_{0t}, \quad \forall t \tag{1.8b}$$

$$Y_t \geq \sum_{i=1}^{k-1} (\tilde{c}_{\phi_t(i)}^t - \tilde{c}_{\phi_t(k)}^t) \gamma_{\phi_t(i)}^t p_{\phi_t(i)}^t + \tilde{c}_{\phi_t(k)}^t \zeta_t + \sum_i (h_t - c_i) p_i^t - h_t D_t - M \mu_{kt}, \quad \forall k, t \tag{1.8c}$$

$$\sum_{k=0}^N \mu_{kt} = 1, \quad \forall t \tag{1.8d}$$

$$Y_t \geq \sum_{i=1}^N \tilde{d}_i^t \eta_i^t p_i^t - \sum_i (s_t + c_i) p_i^t + s_t D_t - M \nu_{0t}, \quad \forall t \tag{1.8e}$$

$$Y_t \geq \sum_{i=1}^{k-1} (\tilde{d}_{\psi_t(i)}^t - \tilde{d}_{\psi_t(k)}^t) \eta_{\psi_t(i)}^t p_{\psi_t(i)}^t + \tilde{d}_{\psi_t(k)}^t \zeta_t - \sum_i (s_t + c_i) p_i^t + s_t D_t - M \nu_{kt}, \quad \forall k, t \tag{1.8f}$$

$$\sum_{k=0}^N \nu_{kt} = 1, \quad \forall t \tag{1.8g}$$

$$\mathbf{p} \in \Omega \tag{1.8h}$$

We will show in the numerical experiment section that it takes substantial amount of time to achieve

a reasonable gap even in small problems. For warm-start purpose, we consider the MINLP formulation in which we directly dualize the second-stage problems in (1.6b) and (1.6c) with extra decision variable $\mathbf{q} \equiv \{\pi_i^t, \lambda_i^t, u_t, \theta_t, \forall i, t\}$.

$$\min_{\mathbf{Y}, \mathbf{p}, \mathbf{q}} \sum_t Y_t \quad (1.9a)$$

$$\text{s.t. } Y_t \geq \sum_{i=1}^N (\gamma_i^t \pi_i^t + h_t - c_i) p_i^t + \zeta_t \mu_t - h_t D_t, \quad \forall t \quad (1.9b)$$

$$Y_t \geq \sum_{i=1}^N (\eta_i^t \lambda_i^t - s_t - c_i) p_i^t + \zeta_t \theta_t + s_t D_t, \quad \forall t \quad (1.9c)$$

$$\pi_i^t + \mu_t \geq \tilde{c}_i^t, \quad \forall i, t \quad (1.9d)$$

$$\lambda_i^t + \theta_t \geq \tilde{d}_i^t, \quad \forall i, t \quad (1.9e)$$

$$\pi_i^t, \mu_t, \lambda_i^t, \theta_t \geq 0, \quad \forall i, t \quad (1.9f)$$

$$\mathbf{p} \in \Omega. \quad (1.9g)$$

1.3.3 Algorithm Framework

To solve the robust demand response scheduling problem (1.4), we first heuristically solve the MINLP problem (1.9) with our proposed algorithm, which converges to a feasible solution. Since the solution \mathbf{p} obtained from problem (1.9) is also feasible for the robust optimization's MIP formulation (1.8), we implement the warm-start solution in the MIP formulation (1.8). We demonstrate that this will numerically speed up the process.

To evaluate the performance of this framework, we consider the following two lowerbounds. The LP relaxation lowerbound within some time limit can serve as a benchmark for performance comparison (MIP gap still cannot close in large numerical examples). In MINLP formulation, we can also consider a lowerbound derived from a convex envelope proposed by McCormick [100].

Off-the-shelf solvers have difficulty solving the MINLP problem directly. We take advantage of the bilinear formulation of the MINLP problem and develop an improved alternating-direction algorithm denoted as ADA to solve the problem (1.9). We will show that iteration of this alternating direction algorithm has a theoretical guarantee of convergence. We also present an improvement of this algorithm.

Alternating Direction Algorithm. We rewrite the MINLP model (1.9) more compactly as

$$\min h(\mathbf{p}, \mathbf{q}) \quad (1.10a)$$

$$\text{s.t. } \mathbf{p} A \mathbf{q}^T + B \mathbf{p} + C \mathbf{q} \leq \mathbf{Y}, \quad (1.10b)$$

$$\mathbf{q} \in \Pi, \quad (1.10c)$$

$$\mathbf{p} \in \Omega, \quad (1.10d)$$

where decision variables are defined as $\mathbf{p} \equiv \{p_i^t\}$ and $\mathbf{q} \equiv \{\pi_i^t, \lambda_i^t, u_t, \theta_t\}$. We also have (1.10a) is defined in (1.9a), (1.10b) corresponds to the bilinear constraints (1.9b) and (1.9c), (1.10c) is the

polyhedron defined by the linear constraints (1.9d)-(1.9f), and (1.10d) is the same as (1.2c)-(1.2h).

When the problem is seperable, it is possible to apply alternating direction algorithm to find heuristic solutions, as discussed by [97]. The main idea of the algorithm is to iteratively fix either \mathbf{p} or \mathbf{q} and search the other direction by solving the remaining problems. The algorithm can be formalized as follows

Algorithm 1 Alternating Direction Algorithm

- 1: **Initialization:** $s = 0$ and $\mathbf{p}_0 \in \Omega$
 - 2: **repeat**
 - 3: Solve $\mathbf{q}_{s+1} = \arg \min h(\mathbf{p}_s, \mathbf{q})$ for $\mathbf{q} \in \Pi$
 - 4: Solve $\mathbf{p}_{s+1} = \arg \min h(\mathbf{p}, \mathbf{q}_{s+1})$ for $\mathbf{p} \in \Omega$
 - 5: $s \leftarrow s + 1$
 - 6: **until** convergence criterion is met.
-

In the following proposition, we summarize that convergence property

Proposition 2. *The sequence of objective function values $\{h(\mathbf{p}_s, \mathbf{q}_s)\}$ generated by Algorithm 2 is convergent.*

We notice that the converged solution may be suboptimal in the global scheme. The convergence we can show via simulation that this returns a very good solution within shorter amount of time. For the initial start point, we use the solution from the deterministic model (1.2).

Improving ADA. In solving for $\mathbf{p}_{s+1} = \arg \min h(\mathbf{p}, \mathbf{q}_{s+1})$, we find out that \mathbf{p}_s from the previous iteration is also a feasible solution in the current iteration. Thus, we can warm start the program with the feasible solution \mathbf{p}_s , which guarantees improvement in each iteration within any time limit.

Meanwhile, when we solve $\mathbf{q}_{s+1} = \arg \min h(\mathbf{p}_s, \mathbf{q})$, we discover that we don't have to solve the entire linear programming problem characterized by objective function (1.9a) and constraints (1.9b) - (1.9f). Instead, we are able to explicitly characterize candidate solutions characterized in Theorem 1. This allows us to simplify the search of \mathbf{q}_{s+1} from solving a large-scale linear programming problem to verifying a small set of candidates that returns the largest objective value. We characterize this property in the following corollary.

Corollary 1. *Given \mathbf{p}_s , optimal solution vector $\mathbf{q}_{s+1}^* = (\pi_i^{s*}, \mu_s^*, \lambda_i^{s*}, \theta_s^*)_{i,t}$ obtained from $\mathbf{q}_{s+1} = \arg \min h(\mathbf{p}_s, \mathbf{q})$ can be obtained by the following enumeration procedures*

Algorithm 2 Improvement in Solving $\mathbf{q}_{s+1} = \arg \min h(\mathbf{p}_s, \mathbf{q})$

- 1: **Iteration:** $t = 1, \dots, T$
 - 2: Enumerate for solution $(\mu_t^*, \pi_1^{t*}, \dots, \pi_N^{t*}) \in \arg \min \{ \sum_{i=1}^N (\gamma_i^t \pi_i^t + h_t - c_i) p_i^t + \zeta_t \mu_t : \mu_t^* \in \{\tilde{c}_1^t, \dots, \tilde{c}_N^t, 0\}, \pi_i^{t*} = (\tilde{c}_i^t - \mu_t^*)^+, \forall i \}$
 - 3: Enumerate for solution $(\theta_t^*, \lambda_1^{t*}, \dots, \lambda_N^{t*}) \in \arg \min \{ (\eta_i^t \lambda_i^t - s_t - c_i) p_i^t + \zeta_t \theta_t : \theta_t^* \in \{\tilde{c}_1^t, \dots, \tilde{c}_N^t, 0\}, \lambda_i^{t*} = (\tilde{c}_i^t - \mu_t^*)^+, \forall i \}$
-

1.4 Computational Experiments

In this section, we study the computation power of the proposed ADA in a large-scale problem environment. We numerically benchmark the heuristic solution with the lowerbound from MIP relaxation and McCormick Envelope. We also record the time it takes for the MIP formulation and the MINLP formulation to achieve the objective value from ADA with the number of resources ranging from 600 to 1200 at different demand levels. In addition, we investigate the solution performance with different reduction error distributions.

1.4.1 Experiment Setup

The DR portfolio in this computation experiments includes three types of DR resources. Type A resource has the most unit profit and the most operational uncertainty, whereas type C resource has the least nominal unit profit and the least operational uncertainty. The unit profit c_i is positively correlated with the capacity p_i^{\max} to reflect the property that DR resources with higher capacity are valued higher by the DR aggregator [65]. Ramping rates and capacity limits are randomly generated within a reasonable range for all DR resources. An example of the resource data is shown in Table 1.1.

Table 1.1: Resource Property

Type	$c_i(\$)$	p_i^{\max}	p_i^{\min}	r_i^+	r_i^-	α_i	β_i
A	25.5	18	4	7	6	-0.5	0.5
B	22.7	15	4	6	5	-0.3	0.3
C	18.5	13	5	6	6	-0.1	0.1

Based on the current industry practice [65], the over-commitment cost h_t is set to be slightly higher than the unit profit because too much supply impairs the economic value of a resource. In contrast, the under-commitments cost s_t is substantially higher than the unit profit because a shortage of commitment can lead to severe penalty from the system operator who suffers from potential power outage. Such penalty includes contractual penalty, damage of reputation, and loss of credibility, which leads to loss of market to competitors and fundamentally hurts the aggregator's DR program. We let the rate of loss depends linearly on the percentage of commitment shortage that the aggregator can bear. For example, we test scenarios where DR aggregator can contain 10% and 30% of commitment shortage. We set the scheduling horizon to be eight periods and one DR event of three periods. We test the performance when the maximum demand level respectively reaches 30% and 80% of the resource total capacity. Table 1.2 shows an example of a DR event with reduction demand of 6000 units.

Table 1.2: Cost Property in a DR Event

t	1	2	3	4	5	6	7	8
$s_t(\$)$	1482.4	1078.8	1485.3	1478.6	1242.7	1400.1	1210.9	1457.9
$h_t(\$)$	34.1	34.5	30.6	33.2	31.4	32.7	34.8	35.2
D_t	0	0	0	6000	6000	6000	0	0

The proposed algorithm is implemented using GUROBI Solver package with 5000 such randomly generated realization samples to evaluate the real profit of the solution.

1.4.2 Computational Analysis

In this section, we discuss the computational performance of ADA in solving the robust DR model (1.4). We record the netagive objective value of the solution obtained by our alternating direction algorithm (row '-Obj'), the time consumption(row 'Time'), the lowerbound from the linear relaxation of MIP formulation and McCormick envelope from the MINLP formulation (row 'MIP Lowerbound' and 'McCormick Lowerbound'), and the optimality gap ('MC Gap' and 'MIP Gap'). The result for $\Gamma_t = 0.01, 0.04, 0.07$ is displayed in table 1.3 .

Table 1.3: Algorithm Performance

$\Gamma_t = 0.01$							
N	200	400	600	800	1000	1200	
-Obj	70541	137560	209530	279190	346596	409104	
Time	81.9	711.2	674.4	1793.5	2189.7	3014.4	
-MIP LB	74381	148040	223499	299028	372559	439205	
MIP Gap	5.16%	7.08%	6.25%	6.63%	6.97%	6.85%	
-McCormick LB	72417	142042	215624	287066	356735	422154	
McCormick Gap	2.59%	3.16%	2.82%	2.74%	2.84%	3.09%	
$\Gamma_t = 0.04$							
N	200	400	600	800	1000	1200	
-Obj	60222	115771	178450	237868	294833	341655	
Time	282.0	487.5	964.6	1619.5	2467.2	4011.1	
-MIP LB	74483	148034	223505	299254	372668	437556	
MIP Gap	19.1%	21.8%	20.2%	20.5%	20.9%	21.9%	
-McCormick LB	69724	136477	207081	275290	342618	404896	
McCormick Gap	13.6%	15.2%	13.8%	13.6%	13.9%	15.6%	
$\Gamma_t = 0.07$							
N	200	400	600	800	1000	1200	
-Obj	50775	95631	149926	199956	246353	279929	
Time	1434.5	828.1	2169.7	3657.5	3736.6	5210.8	
-MIP LB	74492	148071	223522	299227	372706	445821	
MIP Gap	31.8%	35.4%	32.9%	33.2%	32.9%	32.8%	
-McCormick LB	67476	132300	200884	267821	329519	404341	
McCormick Gap	24.7%	27.7%	25.4%	25.3%	25.2%	25.9%	

We observe that for small control level Γ_t , ADA is able to obtain solutions within at most 3% of the optimality gap. As Γ_t increases, the solution quality guarantee deteriorates. We notice that the best performance of the robust model can be achieved at small Γ_t value(ex. $\Gamma_t = 0.01$) in our experiments (see figure 1.2). In general, McCormick envelope formulation provides a better bound for the problem within the numerical time limit. As problem size increases, it is harder to solve the problem.

To make a fair comparison between ADA and solving the MIP formulation directly, we use small problems with only $N = 50$ number of resources at two demand levels. We record the objective value obtained by ADA, and then we also record the time it takes for MINLP and MIP to reach the same objective value. The result is displayed in table 1.4 and 1.5.

Table 1.4: ADA vs MIP in Computation Time with Demand Level at 80%(N = 50)

Γ_t	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
-Obj(\$)	31960	30279	28688	27096	25526	23951	22368	20781	19236	17710
ADA(seconds)	21.2	38.1	35.9	40.3	53.6	35.7	36.6	57.9	41.4	38.2
MIP(seconds)	154.1	886.7	848.3	649.3.9	1212.2	1121.7	1348.7	1224.2	264.2	350.4

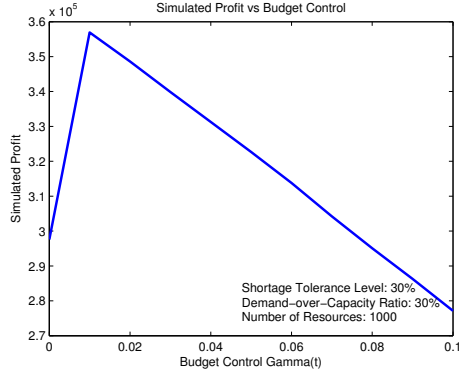
Table 1.5: ADA vs MIP in Computation Time with Demand Level at 30% (N = 50)

Γ_t	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
-Obj(\$)	17634	16728	15844	15006	14184	13394	12595	11794	10993	10195
ADA(seconds)	10.3	18.4	16.4	24.9	23.4	30.2	50.5	69.8	62.6	101.3
MIP(seconds)	87.1	84.9	437.9	486.8	287.2	432.3	535.8	594.2	227.3	216.8

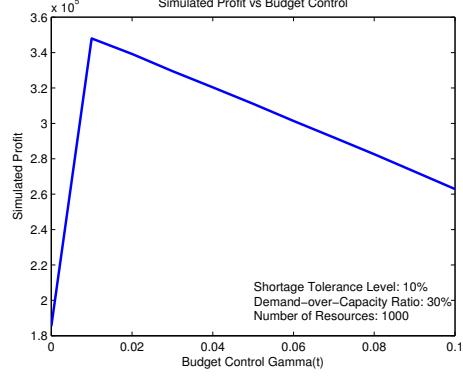
The computation result show that it can take 30 times longer for the branch-and-bound algorithm to get to a solution with the same objective as ADA even in a small problem. The tradeoff is that the MIP formulation has a theoretical guarantee of convergence to global optimality whereas ADA does not.

1.4.3 Profit Analysis

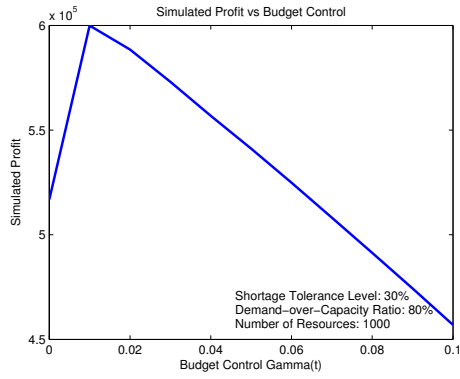
In this section, we study the performance of the robust DR solutions using Monte Carlo simulation. Figure 1.2 shows plots of the average simulation costs of the robust solutions for two total demand reduction levels and two under-commitment cost levels, each with simulation for different uncertainty control levels. Note that $\Gamma_t = 0$ corresponds to the deterministic DR solution.



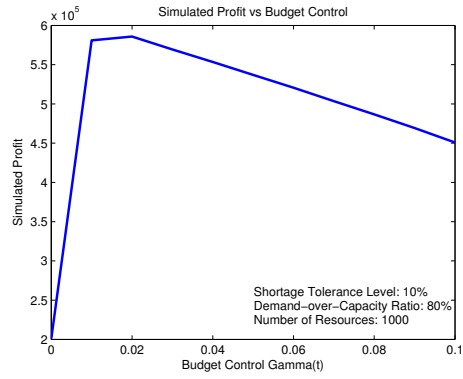
(a) Scenario I: Demand Level 30% and Shortage Tolerance 30%



(b) Scenario II: Demand Level 30% and Shortage Tolerance 10%



(c) Scenario III: Demand Level 80% and Shortage Tolerance 30%



(d) Scenario IV: Demand Level 80% and Shortage Tolerance 10%

Figure 1.2: Reduction Plot

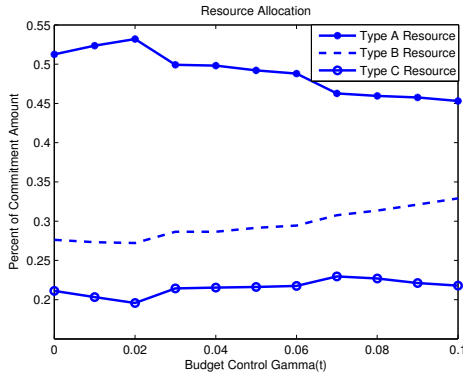
The plots show that when there is a severe penalty from commitment shortage, the robust DR solutions even with a small uncertainty budget Γ_t displays significantly better result from the deterministic solution. We also notice that as the penalty cost increases, the simulated profit from the robust solution decreases much less than from the deterministic solution. The effect comes from the fact that the robust model not only considers the nominal profit of dispatching a resource, but also considers the operational uncertainty from the DR resources. When we increase the demand levels and fix the total resource capacity, the deterministic solution is not able to fully exploit the potential income from demand increase, whereas the robust solution is able to better capitalize the demand market with limited resources.

1.4.4 Solution Analysis

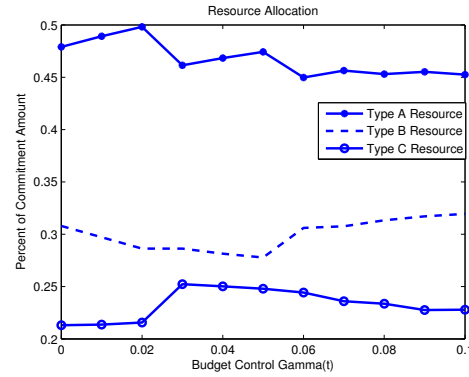
In this section, we discuss the ability of the robust optimization model to avoid high shortage penalty cost in the following two ways

Favoring Resource with less Uncertainty

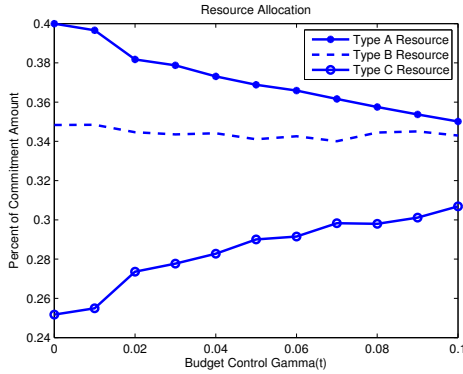
Figure 1.3 shows the percentage of demand reduction amount for each type of resource in the scenario with a total of 1000 resources and uniform distribution for operational error. We observe that type-A resources are generally favored in the deterministic solution, because of its high nominal unit profit. As the control level Γ_t increases, resources with less uncertainty become more favored in all scenarios. The robust optimization model returns more conservative solutions by committing type-B and type-C resources of less uncertainty. This demonstrates the ability of the robust DR model to quantify the tradeoff between the nominal profit and operations uncertainty.



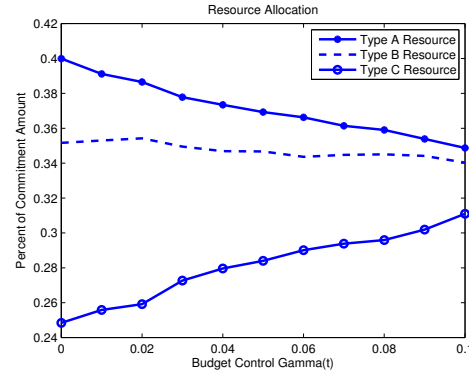
(a) Scenario I: Demand Level 30% and Shortage Tolerance 30%



(b) Scenario II: Demand Level 30% and Shortage Tolerance 10%



(c) Scenario III: Demand Level 80% and Shortage Tolerance 30%



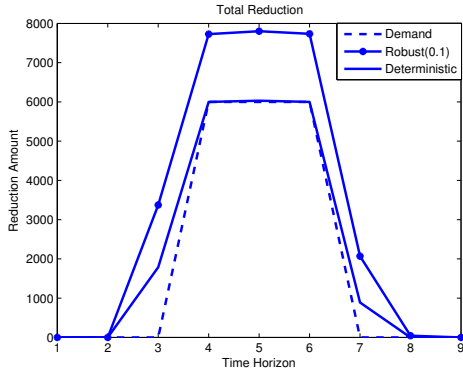
(d) Scenario IV: Demand Level 80% and Shortage Tolerance 10%

Figure 1.3: Allocation Plot

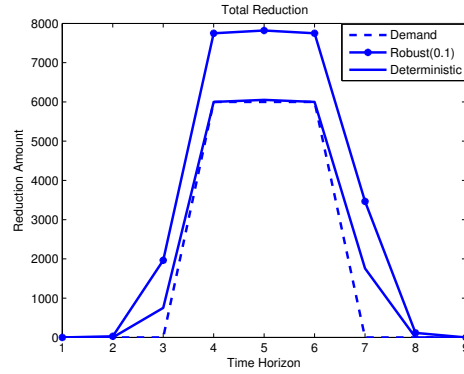
Total Reduction Increases

Since the under-commitment cost is more severe, strategically committing resources to an amount reasonably above the required reduction level substantially reduces the likelihood of under-commitment in actual operations. We observe that the robust DR model is able to justify the cost of effectiveness between under-commitment and over-commitment costs. As shown in Figure 1.4, the total scheduled DR level of the robust solution can be about 2000 units higher in all three demand levels

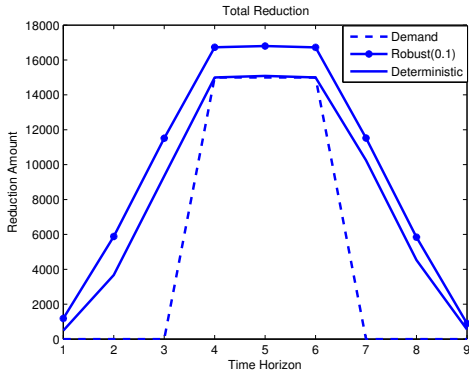
during the peak time.



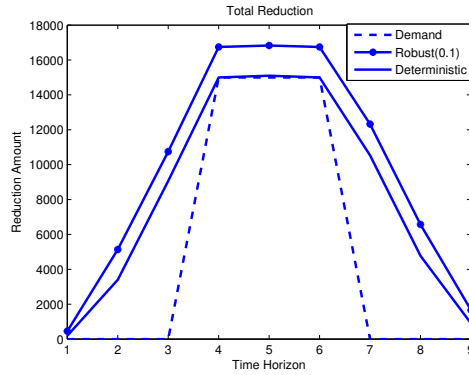
(a) Scenario I: Demand Level 30% and Shortage Tolerance 30%



(b) Scenario II: Demand Level 30% and Shortage Tolerance 10%



(c) Scenario III: Demand Level 80% and Shortage Tolerance 30%



(d) Scenario IV: Demand Level 80% and Shortage Tolerance 10%

Figure 1.4: Allocation Plot

1.5 Conclusions

In this chapter, we explore several important questions on the robust optimization approach for the UC and ED problems that have not been addressed extensively in the existing literature. This includes the question of how to efficiently solve the second-stage problem in the robust UC model, which is a source of significant computational challenge for fully solving these adaptive robust optimization models. We also discuss with examples the properties of the worst case net load scenarios produced by the budget and dynamic uncertainty sets. Many interesting questions are still open, such as solving the multistage robust UC models, designing new dynamic uncertainty sets for solar power and demand response resources, and applying the adaptive robust optimization framework to medium and long term power system planning.

2. Decentralized Algorithms for Solving AC-OPF Using SOCP Relaxation

2.1 Background

AC optimal power flow (OPF) is a basic building block in electric power grid operation. A fundamental question is how to solve AC OPF in a distributed and decentralized fashion; on one hand, multiple ISOs will need to coordinate and jointly optimize generation dispatch over a large geographic area; on the other hand, distributed control of large-number of devices will prevent private information disclosure.

Earlier work in decentralized algorithm for solving AC-OPF problem can be found in [81], in which the authors proposed a regional decomposition approach to divide a large power system into overlapping subsystems. Authors in [137] proposed a fully decentralized algorithm that can decompose the problem into each individual bus, which is one of the first papers applying alternating direction methods of multiplier (ADMM) to the AC OPF problem; however, the algorithm must start with somewhere close to the optimal solutions. Similarly, [50] also decomposed the network into a few sub-regions and solved sub-problems in parallel. Since the problem is non-convex, the proposed algorithm do not have convergence guarantees. Authors in [96] proposed an algorithm that firstly decomposed the problem into smaller sub-problems, then for each non-convex sub-problem, linear approximation and Taylor expansion were used to address the non-convexity issue. [114] applied an ADMM based algorithm to the second-order cone relaxation of the branch-flow model in radial networks. In summary, there are still some major issues remaining: some algorithms may require global coordination; for non-convex problems, the convergence of proposed algorithms in [50, 81, 137] are not guaranteed; semidefinite programming (SDP) relaxation may provide a near-optimal solution, but it is very time-consuming; second-order conic programming (SOCP) relaxation in [114] is only applicable to radial networks.

2.2 Research Approaches

Our objective is to develop decentralized algorithms for approximately solving the AC-OPF problem. The proposed approaches of this project are summarized below:

- Use SOCP as the workhorse for obtaining tight approximation of AC-OPF model
- Use ADMM as the basic algorithmic framework for decentralization
- Study efficient decomposition schemes for constraint and variable splitting
- Study algorithm performance and numerical stability
- Study methods to recover primal variables

2.2.1 SOCP Relaxation of AC-OPF

What is SOCP Relaxation?

The second-order conic programming (SOCP) relaxation of AC-OPF problem is given below:

$$\text{SOCP} - \text{OPF} : \min \sum_{i \in \mathcal{N}} \mathcal{C}_i(p_i^g) \quad (2.1)$$

$$\text{s.t.} \quad p_i^g - p_i^d = G_{ii}c_{ii} + \sum_{j \in \delta(i)} (G_{ij}c_{ij} - B_{ij}s_{ij}) \quad \forall i \in \mathcal{N} \quad (2.2)$$

$$q_i^g - q_i^d = -B_{ii}c_{ii} + \sum_{j \in \delta(i)} (-B_{ij}c_{ij} - G_{ij}s_{ij}) \quad \forall i \in \mathcal{N} \quad (2.3)$$

$$c_{ij}^2 + s_{ij}^2 \leq c_{ii}c_{jj} \quad \forall (i, j) \in \mathcal{L} \quad (2.4)$$

$$\underline{v}_i^2 \leq c_{ii} \leq \overline{v}_i^2 \quad \forall i \in \mathcal{N} \quad (2.5)$$

$$\underline{p}_i^g \leq p_i^g \leq \overline{p}_i^g, \quad \underline{q}_i^g \leq q_i^g \leq \overline{q}_i^g \quad \forall i \in \mathcal{N} \quad (2.6)$$

$$c_{ij} = c_{ji} \quad \forall (i, j) \in \mathcal{L} \quad (2.7)$$

$$s_{ij} = -s_{ji} \quad \forall (i, j) \in \mathcal{L}. \quad (2.8)$$

In the above formulation, \mathcal{N} is the set of buses and \mathcal{L} is the set of transmission branches. Our variables are $p_i^g, q_i^g, c_{ii}, c_{ij}$ and s_{ij} , where p_i^g and q_i^g are the real and reactive power injection at each bus i ; $c_{ij} = e_i e_j + f_i f_j$ and $s_{ij} = e_i f_j - e_j f_i$ for each branch $(i, j) \in \mathcal{L}$, e_i and f_i are the real and imaginary part of the voltage at each bus i , respectively. The objective (2.1) represents the total real generation cost. Constraints (2.2)-(2.3) represent the real and reactive power flow equations; constraint (4.1) relaxes the non-convex relation $c_{ij}^2 + s_{ij}^2 = c_{ii}c_{jj}$ to an SOCP convex constraint. Constraint (2.5) represents the bounds on voltage magnitude, and (2.6) represents the bounds on real and reactive power injection. In summary, we replace bilinear terms in the non-convex power flow equations by single variables c_{ii}, c_{ij}, s_{ij} , so that the power flow constraints become linear. Then non-convexity of power flow equations is relaxed in constraint (4.1).

Why SOCP Relaxation?

The reason that we choose to use SOCP relaxation of AC-OPF is two-fold.

1. Firstly, SOCP relaxation of AC-OPF problem is highly accurate. As we can tell from Figure 2.1, the second column records the objective values (in dollars) for different IEEE Power system test cases (14-bus, 118-bus, 300-bus, and 2869-bus) provided in the MATPOWER library. The Matpower Interior Point Solver (MIPS) is used to get a near-optimal solution for the non-convex AC-OPF problem. Therefore, column two represents feasible solutions that provide an upper bound on the global optimal value. We also coded the SOCP relaxation of AC-OPF problem in Python and used Gurobi solver to get the optimal solution for the convex relaxation problem. As shown in column three, the convex relaxation provides a lower bound on the global optimal value. The small relaxation gap in the last column indicates that our SOCP relaxation is very close to optimal solution and therefore highly accurate.

Case	Matpower obj.val (MIPS)	SOCP obj.val	Relax.Gap
14	8081.52	8074.01	0.09%
118	129660.70	129358.36	0.23%
300	719725.11	718819.59	0.13%
2869	133999.29	133866.62	0.10%

Figure 2.1: Accuracy of SOCP Relaxation for IEEE Test Instances.

2. Secondly, SOCP relaxation is usually much faster than SDP relaxation. As shown in Figure 2.2, the average running time of SOCP for IEEE test instances from 3-bus to 3375-bus is 2.62 seconds, while SDP relaxation needs 380.37 seconds on average. Although generally SDP relaxation is able to provide a tighter lower bound, the solution of SOCP relaxation is more accessible with acceptable quality, as indicated in the third column of Figure 2.2.

	Time (s)	Gap (%)
SOCP	2.62	0.43
SDP	380.37	0.04

Figure 2.2: Comparison of SOCP and SDP Relaxations for IEEE Test Instances up to 3357-Bus System

2.2.2 ADMM

What is ADMM?

ADMM stands for **Alternating Direction Method of Multiplier**, which partitions primal variables into two groups and updates each group alternatively by solving the augmented Lagrangian problem. The problem to be solved is of the form:

$$(P) \quad \min \quad f(x) + g(z) \quad (2.9)$$

$$\text{s.t.} \quad Ax + Bz = c. \quad (2.10)$$

The augmented Lagrangian function can be formulated as

$$L_\rho(x, z, y) = f(x) + g(z) + y^\top (Ax + Bz) + \frac{\rho}{2} \|Ax + Bz - c\|^2. \quad (2.11)$$

ADMM algorithm alternatively minimizes over x and z and updates the dual variable y , using the following iterations

$$\mathbf{x}\text{-update} \quad x^{k+1} = \arg \min_x f(x) + (y^k)^\top (Ax + Bz^k - c) + \frac{\rho}{2} \|Ax + Bz^k - c\|^2 \quad (2.12)$$

$$\textbf{z-update } z^{k+1} = \underset{z}{\arg \min} g(z) + (y^k)^\top (Ax^{k+1} + Bz - c) + \frac{\rho}{2} \|Ax^{k+1} + Bz - c\|^2 \quad (2.13)$$

$$\textbf{dual-update } y^{k+1} = (y^k)^T + \rho(Ax^{k+1} + Bz^{k+1} - c). \quad (2.14)$$

Since we partition the primal variables into two groups: x and z , (2.12) will firstly update x variable by solving the augmented Lagrangian problem in variable x , whose solution is used in (2.13) to update z variable in a similar way; finally, solutions from both (2.12) and (2.13) are used to update the dual variable y .

Why ADMM?

ADMM has deep root in operator splitting and augmented Lagrangian theory. Under mild conditions, the convergence of the ADMM algorithm is guaranteed in the following theorem.[17]

Theorem 2. *If f and g are closed proper convex functions and (2.11) has a saddle point, then, for any $\rho > 0$, ADMM converges with*

- $\|Ax^k - b\| \rightarrow 0$ as $k \rightarrow \infty$
- $f(x^k) + g(z^k) \rightarrow f(x^*) + g(z^*)$ as $k \rightarrow \infty$
- $y^k \rightarrow y^*$ as $k \rightarrow \infty$

where (x^*, z^*) are primal optimal solutions and y^* is dual optimal solution.

Under proper duplication of variables, the x-update and z-update allow a decomposable structure; in other words, the original problem can be reformulated into several subproblems, which can be solved in parallel. Notice that any other constraints (linear, nonlinear) in variable x should be expressed as an indicator function $\mathbf{1}_\Omega(x)$ and added to $f(x)$, while any other constraints involving variable z should be expressed as an indicator function $\mathbf{1}_\Omega(z)$ and added to $g(z)$. An important observation is that the indicator function cannot be coupled, i.e., we cannot have something like $\mathbf{1}_\Omega(x, z)$, which may cause ADMM to diverge. The only constraint that allows coupling of variable should be in the form (10). Details about how to duplicate and split variables will be provided in the next section. The convergence robustness and flexibility in decomposition schemes motivate us to adopt ADMM in this project.

2.2.3 Design of Decomposition Scheme

To begin with, consider a general network flow problem on a given graph $G(V, E)$

$$\begin{aligned} (\mathbf{P}_0) : \text{minimize} \quad & \sum_{i \in V} C_i(n_i) \quad [\text{cost occurs at nodes}] \\ \text{subject to} \quad & [\text{Nodal constraints involving nodal and edge variables:}] \\ & f_i(n_i, (l_{ij})_{j \in \delta(i)}) = 0 \quad \forall i \in V \end{aligned}$$

[Edge constraints involving nodal and edge variables:]

$$g_{ij}(n_i, n_j, l_{ij}) \leq 0 \quad \forall (i, j) \in E.$$

Figure 2.3 shows a simple example on a 2-node graph. In the context of the above formulation, each node has its own nodal variable $n_i, i = 1, 2$, and the line has an edge variable l_e . In addition, we have nodal constraints at each node $f_i, i = 1, 2$, which involves its own nodal variable n_i and the line variable that is incident to this node l_e . Finally, we have an edge constraint g_e , which involves its own edge variable l_e as well as nodal variables at its two endpoints. if we are given a large graph G , we will be working with a huge number of control variables; the interaction between nodal and line variables can be complicated and dynamic since they are coupled together in constraints.

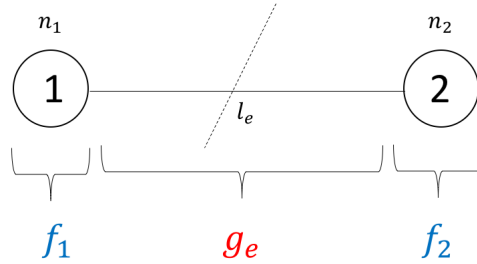


Figure 2.3: Illustration on a 2-Node Graph

Based on our observation, in order to decompose the problem into sub-problems, constraints should be completely separated into each sub-problem. In addition, when solving each sub-problem, we cannot directly use variables from other problems; instead, local variables and local copies of neighboring variables should be used. As a result, it is necessary to **(1)**: assign constraints into sub-problems and **(2)**: duplicate variables so that constraints within each sub-problem do not involve variables from other sub-problems.

We use the 2-node example to illustrate this idea. Consider the problem

$$\min \quad C_1(n_1) + C_2(n_2) \quad (2.15)$$

$$\text{s.t.} \quad f_1(n_1, l_e) = 0 \quad (2.16)$$

$$f_2(n_2, l_e) = 0 \quad (2.17)$$

$$g_e(l_e, n_1, n_2) \leq 0. \quad (2.18)$$

Each node has a nodal constraint $f_i, i = 1, 2$, and there is an edge constraint g_e on the line connecting two nodes. Assume we need to solve the problem in a decentralized manner and have node 1 and 2 to implement independent computation. Clearly, f_1 has to be assigned to node 1 and f_2 has to be assigned to node 2. For the line constraints g_e , there are basically three ways to deal with it.

Firstly we can choose to solve g_e independently, meaning we can separate all three constraints and solve them disjointly in three different smaller optimization problems. Secondly, we can assign edge constraint g_e to one of the endpoint and solve a nodal constraint and the edge constraints together; in other words, we will need to solve two sub-problems, one will have constraints f_i and g_e , while the other one will have the remaining nodal constraint. Thirdly, we can assign g_e to one of the node i , but within the nodal problem of node i , we can again solve f_i and g_e separately in two different sub-subproblems. This situation is similar to the first one but the edge constraint is controlled by one of the node. In terms of computation, these two constraints f_i, g_e can be taken cared of in two different problems. In addition, each of the three situations can be implemented alternatively or in parallel. A summary of the splitting schemes is provided in Figure 2.4.

	X-update	Z-update	Num. duplicated variables	Comments
(a)	$\{f_1, g_e\}$	$\{f_2\}$	6	Solve one nodal and edge constraints jointly, then solve the other nodal constraint
(b)	$\{f_2, g_e\}$	$\{f_1\}$		
(c)	$\{f_1\}$ $\{g_e\}$	$\{f_2\}$	5	Solve one nodal and edge constraints in parallel, then solve the other nodal constraint
(d)	$\{f_2\}$ $\{g_e\}$	$\{f_1\}$		
(e)	$\{f_1\}$ $\{f_2\}$	$\{g_e\}$	4	Solve two nodal problems in parallel, then update edge constraint
(f)	$\{f_1, g_e\}$ $\{f_2\}$	Projection	4	Assign edge to one node, then solve two nodal problems in parallel
(g)	$\{f_2, g_e\}$ $\{f_1\}$	Projection		
(h)	$\{f_1\}$ $\{f_2\}$ $\{g_e\}$	Projection	7	Solve all three constraints in parallel

Figure 2.4: Variable Duplication and Constraint Splitting

(e) and (h) correspond to the first case; (a), (b), (f) and (g) correspond to the second case; (c) and (d) represent the third case. The implementation of each method is summarized in the last column. We adopt (e) for our formulation in the next section since it only requires compact variable duplication and two nodal problems can be solved in parallel.

2.2.4 Decentralized Algorithm

Decomposed SOCP Model

The principles mentioned in previous section lead us to the following variable splitting technique. Each node requires information of its connecting lines, and each line also requires information of its two endpoints. Therefore, we duplicate nodal and line variables so that each node keeps a copy of line variables connected to itself, and each line keeps a copy of its endpoints' variables. One can understand such duplication as the following: each node has some estimation of its adjacent line parameters and, similarly, each line has its own estimation of status of the two endpoints.

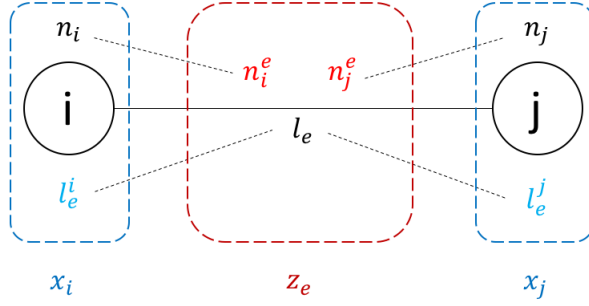


Figure 2.5: Decoupling of Variables

In the context of AC-OPF SOCP formulation, at each node, $n_i = [c_{ii}, (p_i^g), (q_i^g)]$ and at each branch, $l_{ij} = [c_{ij}, s_{ij}]$. Upon duplicating variable, constraints (2.2), (2.3), (2.5)-(2.8) only use x variables and constraint (4.1) only uses z variables. Now constraints can be separated into two groups, and we are ready to implement ADMM to our formulation (**SOCP-OPF**). We adopt the following indices and variables:

- Bus: $i = 1, \dots, N$
- Edge: $(i, j) \in E$ (Assume each line is uniquely represented as (i, j) where $i < j$)
- TieLine: $TL = \{(i, j) | i < j\} \subseteq E$
- Partition: $R_k : k = 1, \dots, K$ (Assume we have a predetermined partition of the whole network)
- c_{ij}^k, s_{ij}^k : region k 's estimation of the line parameters c_{ij} and s_{ij}
- c_i^{ij} : branch (i, j) 's estimation of the variables c_{ii}
- $\delta(i)$: neighbors of bus i
- Iteration index: $r = 1, \dots, \infty$

Using the technique mentioned, (**SOCP-OPF**) can be reformulated as

$$\min \sum_{k=1}^K \sum_{i \in R_k} \mathcal{C}_i(p_i^G) \quad (2.19)$$

s.t. $\forall k = 1, \dots, K :$

$$\begin{aligned} p_i^g - p_i^d &= \\ G_{ii}c_{ii} + \sum_{j \in \delta(i) \cap R_k} (G_{ij}c_{ij} - B_{ij}s_{ij}) + \sum_{j \in \delta(i) \setminus R_k} (G_{ij}c_{ij}^k - B_{ij}s_{ij}^k) &\quad \forall i \in R_k \\ q_i^g - q_i^d &= \end{aligned} \quad (2.20)$$

$$-B_{ii}c_{ii} + \sum_{j \in \delta(i) \cap R_k} (-B_{ij}c_{ij} - G_{ij}s_{ij}) + \sum_{j \in \delta(i) \setminus R_k} (-B_{ij}c_{ij}^k - G_{ij}s_{ij}^k) \quad \forall i \in R_k \quad (2.21)$$

$$\underline{v}_i^2 \leq c_{ii} \leq \overline{v}_i^2, \quad \underline{p}_i^g \leq p_i^g \leq \overline{p}_i^g, \quad \underline{q}_i^g \leq q_i^g \leq \overline{q}_i^g \quad \forall i \in R_k \quad (2.22)$$

$$c_{ij}^2 + s_{ij}^2 \leq c_{ii}c_{jj} \quad \forall (i, j) \in E_k, i, j \in R_k \quad (2.23)$$

$$c_{ij} = c_{ji}, \quad s_{ij} = -s_{ji} \quad \forall (i, j) \in E_k, i, j \in R_k \quad (2.24)$$

$$c_{ij}^k = c_{ji}^k, \quad s_{ij}^k = -s_{ji}^k \quad \forall (i, j) \in TL, i \in R_k \oplus j \in R_k \quad (2.25)$$

$$\begin{aligned} \forall (i, j) \in TL : \\ c_{ij}^2 + s_{ij}^2 \leq c_i^{ij} c_j^{ij} \end{aligned} \quad (2.26)$$

$$\begin{aligned} \forall (i, j) \in TL : \\ i \in R_k, \quad j \in R_h, \quad i < j : \end{aligned}$$

$$c_{ii} = c_i^{ij}, \quad c_{jj} = c_j^{ij} \quad (\theta_1^{ij}, \theta_2^{ij}) \quad (2.27)$$

$$c_{ij}^k = c_{ij}, \quad c_{ij}^h = c_{ij} \quad (\lambda_1^{ij}, \lambda_2^{ij}) \quad (2.28)$$

$$s_{ij}^k = s_{ij}, \quad s_{ij}^h = s_{ij} \quad (\mu_1^{ij}, \mu_2^{ij}). \quad (2.29)$$

Notice that constraints (2.20)-(2.25) can be separated to each sub-region k ; each region is going to solve a AC OPF problem of smaller sizes including these constraints. Since the information about c_{ij} and s_{ij} on tie-lines are not available, each region k will use its estimation c_{ij}^k and s_{ij}^k . Constraint (2.26) can be separated to each tie-line; similarly, since nodal variable c_{ii} and c_{jj} are not available at each tie-line, c_i^{ij} and c_j^{ij} are used instead. In addition, the only coupling constraints(2.27)-(2.29) are linear, which are necessary to enforce consistency between variables and their copies. θ , λ , and μ are the corresponding dual variables for the coupling constraints.

ADMM on Decomposed SOCP Model

Now we can split the variables and constraints for ADMM. The x^k variable for each region R_k includes all the generation variables p_i^g, q_i^g for all generators i in region R_k , all c_{ii} for bus i in R_k , all c_{ij}, s_{ij} for internal line (i, j) in R_k , and all duplicated variables c_{ij}^k, s_{ij}^k for tie line (i, j) between R_k and its neighbor regions. The z variable for each tie line (i, j) between two regions includes c_{ij}, s_{ij} for this line and c_i^{ij}, c_j^{ij} the copy of c_{ii} and c_{jj} kept by the tie line.

In the x-update, we ignore constraint (2.26) on each tie-line. The SOCP model decouples into regional subproblems (we do not include the superscript r of dual variables for simplicity purpose). In particular, each region R_k 's subproblem involves the following variable

$$x^k = ((p_i^g, q_i^g, c_{ii})_{i \in R_k}, (c_{ij}, s_{ij})_{(i,j) \in E_k}, (c_{ij}^k, s_{ij}^k)_{(i,j) \in TL \cap R_k}),$$

and is given below

$$\min \sum_{i \in R_k} C_i(p_i^g) +$$

$$\begin{aligned}
& \sum_{(i,j) \in TL, i \in R_k} \theta_1^{ij} (c_{ii} - (c_i^{ij})^r) + \lambda_1^{ij} (c_{ij}^k - (c_{ij})^r) + \mu_1^{ij} (s_{ij}^k - (s_{ij})^r) + \\
& \sum_{(i,j) \in TL, j \in R_k} \theta_2^{ij} (c_{jj} - (c_j^{ij})^r) + \lambda_2^{ij} (c_{ij}^k - (c_{ij})^r) + \mu_2^{ij} (s_{ij}^k - (s_{ij})^r) + \\
& \frac{\rho}{2} \sum_{(i,j) \in TL, i \in R_k} [(c_{ii} - (c_i^{ij})^r)^2 + (c_{ij}^k - (c_{ij})^r)^2 + (s_{ij}^k - (s_{ij})^r)^2] + \\
& \sum_{(i,j) \in TL, j \in R_k} [(c_{jj} - (c_j^{ij})^r)^2 + (c_{ij}^k - (c_{ij})^r)^2 + (s_{ij}^k - (s_{ij})^r)^2]
\end{aligned} \tag{2.30}$$

$$\begin{aligned}
\text{s.t. } p_i^g - p_i^d = & \sum_{j \in \delta(i) \cap R_k} (G_{ij} c_{ij} - B_{ij} s_{ij}) + \sum_{j \in \delta(i) \setminus R_k} (G_{ij} c_{ij}^k - B_{ij} s_{ij}^k) \quad \forall i \in R_k
\end{aligned} \tag{2.31}$$

$$\begin{aligned}
q_i^g - q_i^d = & -B_{ii} c_{ii} + \sum_{j \in \delta(i) \cap R_k} (-B_{ij} c_{ij} - G_{ij} s_{ij}) + \sum_{j \in \delta(i) \setminus R_k} (-B_{ij} c_{ij}^k - G_{ij} s_{ij}^k) \quad \forall i \in R_k
\end{aligned} \tag{2.32}$$

$$\underline{v}_i^2 \leq c_{ii} \leq \overline{v}_i^2, \quad \underline{p}_i^g \leq p_i^g \leq \overline{p}_i^g, \quad \underline{q}_i^g \leq q_i^g \leq \overline{q}_i^g \quad \forall i \in R_k \tag{2.33}$$

$$c_{ij}^2 + s_{ij}^2 \leq c_{ii} c_{jj} \quad \forall (i, j) \in E_k, i, j \in R_k \tag{2.34}$$

$$c_{ij} = c_{ji}, \quad s_{ij} = -s_{ji} \quad \forall (i, j) \in E_k, i, j \in R_k \tag{2.35}$$

$$c_{ij}^k = c_{ji}^k, \quad s_{ij}^k = -s_{ji}^k \quad \forall (i, j) \in TL, i \in R_k \oplus j \in R_k. \tag{2.36}$$

The x-update is decomposable to each sub-region and they can implement parallel computing. Each sub-region receives dual information (θ, λ, μ) and tie-line's information $((c_i^{ij})^r, (c_{ij})^r, (s_{ij})^r)$, and then solves its own problem. Upon finishing this step, each sub-region can pass its solution to connecting tie-lines and start z-update.

$$\begin{aligned}
\min & \sum_{k=1}^K \sum_{(i,j) \in TL, i \in R_k} \theta_1^{ij} ((c_{ii})^r - c_i^{ij}) + \lambda_1^{ij} ((c_{ij}^k)^r - c_{ij}) + \mu_1^{ij} ((s_{ij}^k)^r - s_{ij}) + \\
& \sum_{k=1}^K \sum_{(i,j) \in TL, j \in R_k} \theta_2^{ij} ((c_{jj})^r - c_j^{ij}) + \lambda_2^{ij} ((c_{ij}^k)^r - c_{ij}) + \mu_2^{ij} ((s_{ij}^k)^r - s_{ij}) + \\
& \frac{\rho}{2} \sum_{k=1}^K \sum_{(i,j) \in TL, i \in R_k} [((c_{ii})^r - c_i^{ij})^2 + ((c_{ij}^k)^r - c_{ij})^2 + ((s_{ij}^k)^r - s_{ij})^2] + \\
& \sum_{k=1}^K \sum_{(i,j) \in TL, j \in R_k} [((c_{jj})^r - c_j^{ij})^2 + ((c_{ij}^k)^r - c_{ij})^2 + ((s_{ij}^k)^r - s_{ij})^2]
\end{aligned} \tag{2.37}$$

$$\begin{aligned}
= & \sum_{(i,j) \in TL, i \in R_k, j \in R_h} \theta_1^{ij}((c_{ii})^r - c_i^{ij}) + \theta_2^{ij}((c_{jj})^r - c_j^{ij}) + \\
& \lambda_1^{ij}((c_{ij}^k)^r - c_{ij}) + \lambda_2^{ij}((c_{ij}^h)^r - c_{ij}) + \\
& \mu_1^{ij}((s_{ij}^k)^r - s_{ij}) + \mu_2^{ij}((s_{ij}^h)^r - s_{ij}) + \\
& \frac{\rho}{2} \{ ((c_{ii})^r - c_i^{ij})^2 + ((c_{jj})^r - c_j^{ij})^2 + \\
& ((c_{ij}^k)^r - c_{ij})^2 + ((c_{ij}^h)^r - c_{ij})^2 + \\
& ((s_{ij}^k)^r - s_{ij})^2 + ((s_{ij}^h)^r - s_{ij})^2 \}
\end{aligned} \tag{2.38}$$

$$\begin{aligned}
\text{s.t. } \forall (i, j) \in TL : \\
c_{ij}^2 + s_{ij}^2 \leq c_i^{ij} c_j^{ij}.
\end{aligned} \tag{2.39}$$

Notice that z-update is decomposable to each tie-line. There is only one constraint on each tie-line, which is second-order cone representable. Actually, z-update is equivalent to solving a projection problem onto a rotated second-order cone in \mathbb{R}^4 , which should be easy and fast to solve. After each tie-line solve its own problem using information from x-update $((c_{ii})^r, (c_{ij}^k)^r, (s_{ij}^k)^r)$, the solution is again passed back to the two neighboring sub-regions.

Finally, the dual update can be decomposed to each tie-line using information from both x-update and z-update. In reality, the dual update can be assigned to either one of the two endpoints of that tie-line.

$$\forall (i, j) \in TL, i \in R_k, j \in R_h :$$

$$(\theta_1^{ij})^{r+1} = (\theta_1^{ij})^r + \rho((c_{ii})^{r+1} - (c_i^{ij})^{r+1}) \tag{2.40}$$

$$(\theta_2^{ij})^{r+1} = (\theta_2^{ij})^r + \rho((c_{jj})^{r+1} - (c_j^{ij})^{r+1}) \tag{2.41}$$

$$(\lambda_1^{ij})^{r+1} = (\lambda_1^{ij})^r + \rho((c_{ij}^h)^{r+1} - (c_{ij})^{r+1}) \tag{2.42}$$

$$(\lambda_2^{ij})^{r+1} = (\lambda_2^{ij})^r + \rho((c_{ij}^k)^{r+1} - (c_{ij})^{r+1}) \tag{2.43}$$

$$(\mu_1^{ij})^{r+1} = (\mu_1^{ij})^r + \rho((s_{ij}^h)^{r+1} - (s_{ij})^{r+1}) \tag{2.44}$$

$$(\mu_2^{ij})^{r+1} = (\mu_2^{ij})^r + \rho((s_{ij}^k)^{r+1} - (s_{ij})^{r+1}) \tag{2.45}$$

After each tie-line finishes its dual update, dual information is passed to connecting sub-regions to start the next iterations. Therefore, in our proposed algorithm, only local communication is

required.

2.3 Numerical Experiments

2.3.1 Effect of Different Partitions on ADMM Convergence

We ran experiments on the proposed algorithm using data from IEEE Power System Test Archive. Cases 14, 30, 118, 300, 1354, pegasus, 2869, pegasus are used. We firstly used an Integer Programming model to partition the graph into several sub-regions that have approximately same size; meanwhile, we minimized the total number of tie-lines. The reasoning is that we believe number of tie-lines connecting two different regions should be controlled. Here we summarized some of them.

Case	Num. of Partitions	Num. of Tie-lines	ρ	Num. Iter	Obj. Gap	Serial Time (sec)	Est. Parallel Time (sec)
14	2	3	100	96	0.0390%	32.7624	16.3812
	3	5	400	50	0.0675%	27.1084	9.0361
	4	6	600	50	0.0342%	33.5219	8.3805
118	2	4	200	28	0.4286%	13.1334	6.5667
	3	7	600	34	1.3969%	24.1909	8.0636
	4	11	1600	52	0.6356%	62.3426	15.5856
300	2	4	1600	93	0.0002%	60.3942	30.1971
	3	7	1600	94	0.0458%	93.1572	31.0524
	4	8	1000	134	0.4199%	127.5870	31.8967
2869	2	6	100	158	0.0006%	340.1150	170.0575
	3	12	100	148	0.0015%	344.4219	114.8073
	4	19	300	103	0.0044%	238.3519	59.5880

Figure 2.6: Numerical Result of Proposed Algorithm

We partitioned the network into 2, 3 and 4 sub-regions of similar sizes, and the total number of tie-lines are in the third column. Then we tried different values of penalty parameter ρ to compare convergence speed in terms of number of iterations needed, as shown in column 4 and 5. Column 6 records the objective gap with the centralized SOCP relaxation solution. As we can tell from the table, with properly chosen ρ , the algorithm can converge in a few tens of iterations and the objective is acceptable. In addition, the behavior of the algorithm is consistent for both small and large systems, so we conclude the algorithm is highly scalable. Since we implemented the ADMM update in serial, we expect the running time should be efficiently reduced assuming each sub-region is able to conduct independent and concurrent computation. The last two columns represent the time in serial and in parallel.

3. A New Voltage Stability-Constrained Optimal Power Flow Model: Sufficient Condition, SOCP Representation, and Relaxation

3.1 Introduction

The need to ensure steady-state voltage stability and maintain sufficient loading margin in optimal power flow (OPF) models has led to the development of voltage stability-constrained OPF (VSC-OPF) models, which solves OPF problems while accounting for voltage stability limits at the same time. Traditionally, to avoid system instability, security constraints such as voltage magnitude limits and line flow limits are enforced in normal OPF models. However, the effectiveness of these security constraints alone in safeguarding system stability may be insufficient in modern power systems with adequate reactive power support, which is demonstrated by a two-bus example in [140]. Another motivation for the inclusion of steady-state stability limit in an OPF formulation is the increasing trend to operate power systems ever closer to their operational limits due to increased demand and competitive electricity market. Without stability constraints, the robustness of the OPF solution against voltage instability is not ensured.

To formulate a VSC-OPF problem, the model in [22] uses two sets of power flow equations representing base loading and critical loading conditions, the power injections of which are related by the loading factor to be optimized. The model is extended to a multi-objective one in [102] in which voltage stability and social welfare are simultaneously taken care of. An extension to incorporate $N - 1$ contingencies has been reported in [103] where a heuristic contingency ranking technique is applied for computation tractability. An alternative method to account for contingencies in a VSC-OPF model based on iterative CPF-OPF computation is presented in [101]. However, the loading margin is only quantified along one direction of power variation in these models. Dynamic voltage stability has been considered in security-constrained OPF such that systems subject to contingencies will settle down to stable operating points. Dynamic simulation with scenario filtering techniques have been employed to this end in [23, 24]. These methods are highly dependent on the choice of contingencies and suffer from scalability issue. A different strategy to represent proximity to voltage instability is through the use of minimum singular value (MSV) of the power flow Jacobian, which can be used as a stability constraint in a VSC-OPF model. The main drawbacks of the method are that 1) the physical meaning of MSV is unclear; 2) MSV is not an explicit function of the optimization variables. Linearization and iterative algorithms have been proposed trying to address the second issue [6, 86]. However, the computational cost is prohibitively high for large-scale systems.

To circumvent the weaknesses of the aforementioned VSC-OPF models and achieve a better trade-off between robustness and computational tractability, several heuristic voltage stability indicators have been embedded in VSC-OPF formulations. For instance, the L -index originally proposed in [80] has been used as an indicator for voltage stability improvement in [88]. Leveraging semidefinite programming (SDP) relaxation of OPF, this problem can be formulated as an SDP with quasi-convex objective [113]. Polyhedron approximation of security boundaries has been applied in a DC-OPF model in [33]. However, the approach does not scale well with the dimension of feasible region. The sufficient voltage stability condition for reactive power flow equations in [132] has been used for voltage stress minimization in [140]. A voltage stability index based on branch flow

is integrated in VSC-OPF formulation in [156]. Major concerns of these indices are their conservativeness and computational properties. Hence, the main motivation of this paper is to apply a novel and tight voltage stability index in the VSC-OPF model which enjoys nice computational properties under very mild approximation.

We first introduce a sufficient condition for power flow Jacobian nonsingularity that we proposed recently in [150]. We then formulate a VSC-OPF problem in which the voltage stability margin is quantified by the condition. We show that when load powers are fixed, this voltage stability condition describes a second-order conic representable set in a transformed voltage space. Thus second-order cone program (SOCP) reformulation can naturally incorporate the condition. Notice that the formulation does not require the DC or decoupled power flow assumptions. To improve computation time, we sparsify the dense stability constraints while preserving very high accuracy.

The rest of the paper is organized as follows. Section 3.2 provides background on power system modeling. The sufficient condition for power flow Jacobian nonsingularity is introduced in Section 3.3. We discuss the VSC-OPF formulation, its convex reformulation, and sparse approximation in Section 3.4. Section 3.5 presents results of extensive computational experiments and comparative studies. Section 3.6 concludes.

3.2 Background

3.2.1 Notations

The cardinality of a set or the absolute value of a (possibly) complex number is denoted by $|\cdot|$. $i = \sqrt{-1}$ is the imaginary unit. \mathbb{R} and \mathbb{C} are the set of real and complex numbers, respectively. For vector $x \in \mathbb{C}^n$, $\|x\|_p$ denotes the p -norm of x where $p \geq 1$ and $\text{diag}(x) \in \mathbb{C}^{n \times n}$ is the associated diagonal matrix. The n -dimensional identity matrix is denoted by \mathbf{I}_n . $\mathbf{0}_{n \times m}$ denotes an $n \times m$ matrix of all 0's. For $A \in \mathbb{C}^{n \times n}$, A^{-1} is the inverse of A . For $B \in \mathbb{C}^{m \times n}$, B^T , B^H are respectively the transpose and conjugate transpose of B , and B^* is the matrix with complex conjugate entries. The real and imaginary parts of B are denoted as $\text{Re } B$ and $\text{Im } B$. b_i denotes the vector formed by the i th row of B .

3.2.2 Power System Modeling

We consider a connected single-phase power system with $n + m$ buses operating in steady-state. The underlying topology of the system can be described by an undirected connected graph $G = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \mathcal{N}_G \cup \mathcal{N}_L$ is the set of buses equipped with (\mathcal{N}_G) and without (\mathcal{N}_L) generators (or generator buses and load buses), and that $|\mathcal{N}_G| = m$ and $|\mathcal{N}_L| = n$. We number the buses such that the set of load buses are $\mathcal{N}_L = \{1, \dots, n\}$ and the set of generator buses are $\mathcal{N}_G = \{n+1, \dots, n+m\}$. Generally, for a complex matrix $A \in \mathbb{C}^{(n+m) \times k}$, define $A_L = (A_{ij})_{i \in \mathcal{N}_L}$. That is, A_L is the first n rows of the matrix A . Similarly, define $A_G = (A_{ij})_{i \in \mathcal{N}_G}$. Every bus i in the system is associated with a voltage phasor $V_i = |V_i|e^{i\theta_i}$ where $|V_i|$ and θ_i are the magnitude and phase angle of the voltage. We will find it convenient to adopt rectangular coordinates for voltages sometimes, so we also define $V_i = e_i + if_i$. The generator buses are modeled as PV buses, while load buses are modeled as PQ buses. For bus i , the injected power is given as $S_i = P_i + iQ_i$.

The line section between buses i and j in the system is weighted by its complex admittance $y_{ij} = 1/z_{ij} = g_{ij} + ib_{ij}$. The nodal admittance matrix $Y = G + iB \in \mathbb{C}^{(n+m) \times (n+m)}$ has components $Y_{ij} = -y_{ij}$ and $Y_{ii} = y_{ii} + \sum_{j=1}^{n+m} y_{ij}$ where y_{ii} is the shunt admittance at bus i .

The nodal admittance matrix relates system voltages and currents as

$$\begin{bmatrix} I_L \\ I_G \end{bmatrix} = \begin{bmatrix} Y_{LL} & Y_{LG} \\ Y_{GL} & Y_{GG} \end{bmatrix} \begin{bmatrix} V_L \\ V_G \end{bmatrix}. \quad (3.1)$$

We obtain from (3.1) that

$$V_L = -Y_{LL}^{-1}Y_{LG}V_G + Y_{LL}^{-1}I_L. \quad (3.2)$$

Define the vector of equivalent voltage to be $E = -Y_{LL}^{-1}Y_{LG}V_G$ and the impedance matrix to be $Z = Y_{LL}^{-1}$ (we assume the invertibility of Y_{LL} and note that this is almost always the case for practical systems). With the definitions, (3.2) can be rewritten as

$$V_L = E + ZI_L. \quad (3.3)$$

For practical power systems, the generator buses have regulated voltage magnitudes and small phase angles. It is common in voltage stability analysis to assume that the generator buses have constant voltage phasor V_G [80, 150]. The assumption can be partially justified by the fact that voltage instability are mostly caused by system overloading due to excess demand at load side, irrelevant of generator voltage variations.

Assumption 1. *The vector of generator bus voltages V_G is constant.*

Note that Assumption 1 is always satisfied for uni-directional distribution systems where the only source is modeled as a slack bus with fixed voltage phasor. The voltage stability constraint in the paper is based on our recent result on the nonsingularity of power flow Jacobian [150]. The derivation of the result takes advantage of the special characteristics of systems with constant generator voltage vector E . With Assumption 1, E is fixed and the result in [150] can be applied.

The power flow equations in the rectangular form relate voltages and power injections at each bus $i \in \mathcal{N}$ via

$$P_i = \sum_{j=1}^{n+m} [G_{ij}(e_i e_j + f_i f_j) + B_{ij}(e_j f_i - e_i f_j)], \quad (3.4a)$$

$$Q_i = \sum_{j=1}^{n+m} [G_{ij}(e_j f_i - e_i f_j) - B_{ij}(e_i e_j + f_i f_j)]. \quad (3.4b)$$

Remark 1. The power flow Jacobian with Assumption 1 is given by

$$J_{LL} := \begin{bmatrix} \frac{\partial P_L}{\partial e_L} & \frac{\partial P_L}{\partial f_L} \\ \frac{\partial Q_L}{\partial e_L} & \frac{\partial Q_L}{\partial f_L} \end{bmatrix}. \quad (3.5)$$

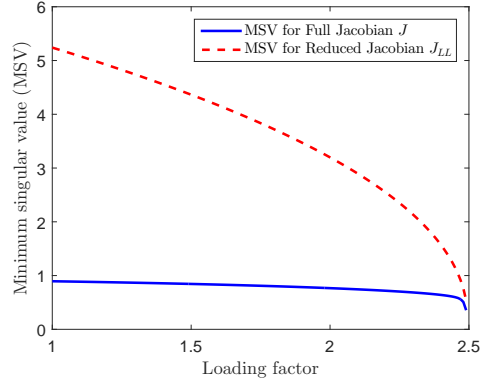


Figure 3.1: MSVs of Full and Reduced Power Flow Jacobian with Respect to System Loading for 9-Bus System

Note that J_{LL} is in fact a submatrix of the full Jacobian considering generator real power equations:

$$J = \begin{bmatrix} J_{GG} & J_{GL} \\ J_{LG} & J_{LL} \end{bmatrix}. \quad (3.6)$$

As we know, voltage stability studies are primarily concerned with the singularity of power flow Jacobian J . Of course, for generic matrix J , the singularity of its principal submatrices are not necessarily related to that of the full matrix. Then the assumption of constant generator voltage phasors seems to be questionable since the stability analysis based on a submatrix may not be relevant. However, we note this is not the case in voltage stability analysis. First of all, the validity of using power flow Jacobian as a voltage stability indicator is based on the assumption that $\det J_{LL} \neq 0$. In this case the system stability is determined by the reduced Jacobian $J_{\text{red}} = J_{GG} - J_{GL}J_{LL}^{-1}J_{LG}$, whose determinant is singular if and only if the determinant of the power flow Jacobian

$$\det J = \det J_{LL} \det(J_{GG} - J_{GL}J_{LL}^{-1}J_{LG}) \quad (3.7)$$

is singular [143, Chap. 5]. However, the singularity of J_{LL} is itself one of the mechanisms of voltage collapse, which is called singularity-induced bifurcation and has been demonstrated through a rudimentary dynamic power system model in [144]. Second, the singularity of J is often associated with the ill-conditioning of the matrix J_{LL} ; and the MSV of J_{LL} tends to decrease monotonically with increased loading levels, which are demonstrated by IEEE 9-bus system in Fig. 3.1. Therefore we believe the study of J_{LL} for voltage stability purposes can be justified from both physical and numerical perspective.

Remark 2. After the overexcitation limiter of a generator takes effect, the terminal voltage of the generator can no longer be regulated, and a common modeling practice is to switch the bus type from PV to PQ. We note that the generator can also be modeled as a constant excitation emf behind synchronous impedance based on [143, Sect. 3.4.2], the validity of which has been justified in [149]. The synchronous impedance can be absorbed by the system admittance matrix and the model reduces to the one with constant voltage sources and constant power load buses. This can be done iteratively every time generators reaches their reactive power limits after OPF computation.

3.2.3 AC-OPF Formulation

Using the power flow equations (3.4a)-(3.4b), a standard AC-OPF model can be written as

$$\min \sum_{i \in \mathcal{N}_G} f_i(P_{G_i}) \quad (3.8a)$$

$$\text{s.t. } P_i(e, f) = P_{G_i} - P_{D_i}, \quad i \in \mathcal{N} \quad (3.8b)$$

$$Q_i(e, f) = Q_{G_i} - Q_{D_i}, \quad i \in \mathcal{N} \quad (3.8c)$$

$$\underline{P}_{G_i} \leq P_{G_i} \leq \overline{P}_{G_i}, \quad i \in \mathcal{N}_G \quad (3.8d)$$

$$\underline{Q}_{G_i} \leq Q_{G_i} \leq \overline{Q}_{G_i}, \quad i \in \mathcal{N}_G \quad (3.8e)$$

$$\underline{V}_i^2 \leq e_i^2 + f_i^2 \leq \overline{V}_i^2, \quad i \in \mathcal{N} \quad (3.8f)$$

$$|P_{ij}(e, f)| \leq \overline{P}_{ij}, \quad (i, j) \in \mathcal{E} \quad (3.8g)$$

$$|I_{ij}(e, f)| \leq \overline{I}_{ij}, \quad (i, j) \in \mathcal{E}, \quad (3.8h)$$

where $f_i(P_{G_i})$ in (3.8a) is the variable production cost of generator i , assuming to be a convex quadratic function; P_{G_i} and P_{D_i} in (3.8b)-(3.8c) are the real power generation and load at bus i , respectively; Q_{G_i} and Q_{D_i} are the reactive power generation and load at bus i ; $P_i(e, f)$ and $Q_i(e, f)$ are given by the power flow equations (3.4); constraints (3.8d)-(3.8e) represent the real and reactive power generation capability of generator i . P_{ij} and I_{ij} in (3.8g)-(3.8h) are the real power and current magnitude flowing from bus i to j for line $(i, j) \in \mathcal{E}$, respectively.

3.3 A Sufficient Condition for Nonsingularity of Power flow Jacobian

A sufficient condition for the nonsingularity of power flow Jacobian is recently proposed in [150] as stated in the following theorem. We will use this result to derive a voltage stability index which is to be embedded in an OPF model to form a VSC-OPF formulation.

Theorem 3. *The power flow Jacobian of (3.4) is nonsingular if*

$$|V_i| - \|z_i^T \text{diag}(I_L)\|_1 > 0, \quad i \in \mathcal{N}_L. \quad (3.9)$$

The proof is based on similarity transformation of the power flow Jacobian. We have shown that the transformed matrix is strictly diagonally dominant as long as (3.9) holds. Since strictly diagonally dominant matrices are nonsingular and similarity transformation preserves eigenvalues, the power flow Jacobian is nonsingular when (3.9) holds. The proof takes advantage of the special structure of the matrix J_{LL} . Under Assumption 1, the power flow Jacobian and J_{LL} coincide. For proof of the theorem, see [150].

The term $\|z_i^T \text{diag}(I_L)\|_1$ in Theorem 1 can be thought of as the generalized voltage drop between the equivalent source with voltage E_i to the load. Then the theorem states that the system is voltage stable if the generalized voltage drop is less than the corresponding load voltage magnitude for all load buses. It has been shown in [150] that the result is strong, meaning that the violation of the condition is often immediately followed by the loss of voltage stability.

It is suggested in [80] that the following condition is satisfied at the point of voltage instability under certain simplifying assumptions (proportional load current variations, etc.)

$$\left| \sum_{i=1}^n Z_{ji} I_i \right| = |V_j|. \quad (3.10)$$

It is seen from (3.3) that the left hand side is the voltage drop between the equivalent source with voltage E_j and the load. The result implies that under certain assumptions, the voltage stability of a multi-bus system resembles that of a two-bus system where the voltage stability boundary is achieved when the magnitude of voltage drop and load voltage are identical. Due to various assumptions, the condition works relatively well under proportional load variations, but becomes less effective as load variation deviates from the assumed proportional pattern.

We note the similarity between the condition (3.10) and (3.9) used in the paper. The condition (3.9) is weaker in the sense that the generalized voltage drop $\|z_i^T \text{diag}(I_L)\|_1$ is larger than the actual voltage drop in (3.10), but it nevertheless generalizes the latter condition and does not require the proportional current injection assumption. For a more thorough comparison of the two conditions, see [150].

3.4 A New Model for VSC-OPF

The standard AC-OPF formulation embeds system security constraints as line real power and current limits in (3.8g) and (3.8h). However, the parameters in these security-related constraints, such as \bar{P}_{ij} and \bar{I}_{ij} , are calculated off-line using possible dispatch scenarios that do not necessarily represent the actual system conditions [102]. This motivates the formulation of VSC-OPF models. In this section, we propose a new model for VSC-OPF using the voltage condition derived in (3.9) and show that it has nice convex properties amenable for efficient computation.

3.4.1 New Formulation

We propose the following new VSC-OPF model,

$$\min \sum_{i \in \mathcal{N}_G} f_i(P_{G_i}) \quad (3.11a)$$

$$\text{s.t.} \quad (3.8b) - (3.8f)$$

$$|V_i| - \sum_{j=1}^n \frac{A_{ij}}{|V_j|} \geq \underline{t}_i, \quad i \in \mathcal{N}_L. \quad (3.11b)$$

where $A_{ij} := |Z_{ij} S_j|$. The key constraint is (3.11b), which reformulates the left-hand side of (3.9) by writing line currents as the ratio of apparent powers that satisfy the power flow equations (3.4) and voltages, and \underline{t}_i is a preset positive parameter to control the level of voltage stability. We note that line flow constraints are not included in the VSC-OPF formulation (3.11). We have deliberately chosen not to include them since 1) we would like to demonstrate the capability of the proposed voltage stability constraint in restraining system margins to voltage instability, and 2) we

believe the proposed constraint is better suited for stability constraining purposes. To guarantee the same level of voltage stability, line flow constraints come at a price of higher level of conservativeness compared to the proposed stability constraint since the line flow constraints are not intrinsic voltage stability measures. It then follows that to ensure similar level of voltage stability, the inclusion of line flow constraints shrinks the feasibility region of the problem. Of course, there are no technical difficulties in the inclusion of line flow constraints in our formulation and we agree that for lines with low thermal ratings or low line flow margins, the inclusion of corresponding constraints are necessary and beneficial. To ensure that (3.11) is a proper formulation with good computational property, we first show that the set of voltages satisfying condition (3.11b) is voltage stable, and then we show that (3.11b) is second-order cone (SOC) representable, thus convex, when S_L is constant. The condition of constant S_L is always met in OPF problems.

Connectedness

A necessary condition for voltage instability is the singularity of power flow Jacobian [143, Sect. 7.1.2]. Assume that the zero injection solution of power flow equations (3.4) is voltage stable with a nonsingular Jacobian (which always holds for any physically meaningful system). We know from (3.4) that every entry of J is a continuous function of voltages, so the eigenvalues of J are also continuous in voltages. Since a continuous function maps a connected set to another connected set, if a given connected set of power flow solutions contains the zero injection solution (which is voltage stable) and the corresponding power flow Jacobian of every point in the set is nonsingular, then the set characterizes a subset of voltage stable solutions. Define the set $\mathcal{S}_0 := \{V_L \mid (3.9) \text{ holds}\}$ and $\mathcal{S}_0 \supseteq \mathcal{S}_t := \{V_L \mid (3.11b) \text{ holds}\}$. We know from Theorem 3 that the power flow Jacobian is nonsingular for $V_L \in \mathcal{S}_0$, we also know the zero injection solution is in \mathcal{S}_0 . Therefore, in order to show the set \mathcal{S}_t is voltage stable, we show the more general case that \mathcal{S}_0 is voltage stable, which amounts to showing the connectedness of \mathcal{S}_0 . We give the proof of this property below.

Theorem 4. *The set \mathcal{S}_0 is connected.*

Proof. To show the set is connected, we fix a point in the set and show that for any other point in the set, the line segment between the two points lies in the set.

When load currents are all zero, it follows from (3.3) that the nodal load voltages are simply E . We denote the zero injection voltage solution by v_0 , that is, $v_0 := E$. Then $v_0 \in \mathcal{S}_0$ follows immediately since $\|z_i^T \text{diag}(I_L)\|_1 = 0$ for all $i \in \mathcal{N}_L$. Take $v_1 \in \mathcal{S}_0$ and define V_L parametrized by $t \in [0, 1]$ as $V_L(t) = v_0 + (v_1 - v_0)t$. We will show $V_L(t)$ is in \mathcal{S}_0 . It is clear that current injections are linear functions of t , since we know from (3.3) that

$$I_L(t) = Y_{LL} (V_L(t) - E) \quad (3.12a)$$

$$= Y_{LL} (v_0 + (v_1 - v_0)t - v_0) \quad (3.12b)$$

$$= Y_{LL} (v_1 - v_0)t. \quad (3.12c)$$

We claim that for any $t \in [0, 1]$ the derivative of $\sum_{j=1}^n |Z_{ij} I_j|$ is larger than or equal to the magnitude of that of $|V_i|$ for all $i \in \mathcal{N}_L$. Since current injections are linear in t , let $Z_{ij} I_j$ be denoted by

$a_{ij}t + ib_{ij}t$ for real numbers a_{ij} and b_{ij} for all $(i, j) \in \mathcal{N}_L \times \mathcal{N}_L$, and denote $a := \sum_{j=1}^n a_{ij}$ and $b := \sum_{j=1}^n b_{ij}$ for brevity, then for each $i \in \mathcal{N}_L$ we have

$$\frac{d}{dt} \left(\sum_{j=1}^n |Z_{ij} I_j| \right) = \frac{d}{dt} \left(\sum_{j=1}^n \sqrt{a_{ij}^2 + b_{ij}^2} \right) t \quad (3.13a)$$

$$= \sum_{j=1}^n \sqrt{a_{ij}^2 + b_{ij}^2} \quad (3.13b)$$

$$\geq \sqrt{a^2 + b^2}, \quad (3.13c)$$

where the inequality is due to successive application of trigonometric inequality. On the other hand, the voltage magnitude $|V_i|$ is

$$\begin{aligned} |V_i| &= |v_{0,i} + z_i^T I_L| \\ &= \sqrt{\left(\operatorname{Re}(v_{0,i}) + \sum_{j=1}^n a_{ij}t \right)^2 + \left(\operatorname{Im}(v_{0,i}) + \sum_{j=1}^n b_{ij}t \right)^2}, \end{aligned} \quad (3.14)$$

and the derivative of $|V_i|$ with respect to t is

$$\frac{d|V_i|}{dt} = \frac{a(\operatorname{Re}(v_{0,i}) + at) + b(\operatorname{Im}(v_{0,i}) + bt)}{\sqrt{(\operatorname{Re}(v_{0,i}) + at)^2 + (\operatorname{Im}(v_{0,i}) + bt)^2}}. \quad (3.15)$$

Then, by Cauchy-Schwarz inequality we have $|d|V_i|/dt| \leq \sqrt{a^2 + b^2}$. Comparing with (3.13), we see the claim holds.

Suppose $\sum_{j=1}^n |Z_{ij} I_j(t_1)| \geq |V_i(t_1)|$ for some $t_1 \in (0, 1)$ and $i \in \mathcal{N}_L$, then based on the fundamental theorem of calculus we have

$$\sum_{j=1}^n |Z_{ij} I_j(1)| = \sum_{j=1}^n |Z_{ij} I_j(t_1)| + \int_{t_1}^1 \left(\sum_{j=1}^n |Z_{ij} I_j| \right)' dt \quad (3.16a)$$

$$\geq \sum_{j=1}^n |Z_{ij} I_j(t_1)| + \sqrt{a^2 + b^2}(1 - t_1), \quad (3.16b)$$

and

$$|V_i(1)| = |V_i(t_1)| + \int_{t_1}^1 |V_i(t)|' dt \quad (3.17a)$$

$$\leq |V_i(t_1)| + \sqrt{a^2 + b^2}(1 - t_1). \quad (3.17b)$$

The two inequalities imply that $\sum_{j=1}^n |Z_{ij} I_j(1)| \geq |V_i(1)|$, which is a contradiction since $v_1 \in \mathcal{S}_0$. We conclude the line segment between v_0 and v_1 lies in \mathcal{S}_0 . \square

SOC Representation of Voltage Stability Constraint

The voltage stability constraint (3.11b) is not directly a convex constraint in the voltage variable V_i , however, we show that it can be reformulated as a convex constraint, more specifically, an SOC constraint in squared voltage magnitude $|V_i|^2$ providing S_L is fixed. This SOC reformulation will be utilized in the following section for SOCP relaxation of VSC-OPF.

Proposition 3. *Constraint (3.11b) is SOC representable in the squared voltage magnitude $|V_i|^2$'s, i.e. (3.11b) can be reformulated using SOC constraints in $|V_i|^2$'s.*

Proof. First of all, introduce variable $c_{ii} := |V_i|^2$, and x_i, z_i for each bus $i \in \mathcal{N}_L$ such that

$$x_i \leq \sqrt{c_{ii}}, \quad (3.18)$$

$$x_i z_i \geq 1, \quad (3.19)$$

$$x_i \geq 0.$$

Note that $x_i z_i = (\frac{x_i + z_i}{2})^2 - (\frac{x_i - z_i}{2})^2$ and $c_{ii} = (\frac{c_{ii} + 1}{2})^2 - (\frac{c_{ii} - 1}{2})^2$, then we see both (3.18) and (3.19) can be rewritten as the following SOC constraints

$$\sqrt{x_i^2 + \frac{(c_{ii} - 1)^2}{4}} \leq \frac{c_{ii} + 1}{2}, \quad (3.20)$$

$$\sqrt{1 + \frac{(x_i - z_i)^2}{4}} \leq \frac{x_i + z_i}{2}. \quad (3.21)$$

Therefore, by defining $A_{ij} = |Z_{ij} S_j|$, (3.11b) can be equivalently represented as

$$x_i - \sum_{j=1}^n A_{ij} z_j \geq t_i, \quad (3.22a)$$

$$\| [x_i, (c_{ii} - 1)/2]^T \|_2 \leq (c_{ii} + 1)/2, \quad (3.22b)$$

$$\| [1, (x_i - z_i)/2]^T \|_2 \leq (x_i + z_i)/2, \quad (3.22c)$$

$$x_i \geq 0, \quad (3.22d)$$

for every bus $i \in \mathcal{N}_L$, which are SOCP constraints. \square

3.4.2 SOCP Relaxation of VSC-OPF

By Proposition 3, the voltage stability condition (3.9) is reformulated as SOCP constraints (3.22). However, the power flow equations (3.8b)-(3.8c) are still nonconvex. In the following, we propose an SOCP relaxation of the proposed VSC-OPF model (3.11) by combining the SOC reformulation of the voltage stability constraint (3.22) with the recent development of SOCP relaxation of standard AC-OPF [85]. In particular, for each line $(i, j) \in \mathcal{E}$, define

$$c_{ij} = e_i e_j + f_i f_j \quad (3.23a)$$

$$s_{ij} = e_i f_j - e_j f_i. \quad (3.23b)$$

An implied constraint of (3.23a)-(3.23b) is the following:

$$c_{ij}^2 + s_{ij}^2 = c_{ii}c_{jj}. \quad (3.24)$$

Now we can introduce the following SOCP relaxation of the VSC-OPF model (3.11) in the new variables c_{ii} , c_{ij} , and s_{ij} as follows

$$\begin{aligned} \min \quad & \sum_{i \in \mathcal{N}_G} f_i(P_{G_i}) \\ \text{s.t.} \quad & P_{G_i} - P_{D_i} = G_{ii}c_{ii} + \sum_{j \in N(i)} P_{ij}, \quad i \in \mathcal{N} \end{aligned} \quad (3.25a)$$

$$Q_{G_i} - Q_{D_i} = -B_{ii}c_{ii} + \sum_{j \in N(i)} Q_{ij}, \quad i \in \mathcal{N} \quad (3.25b)$$

$$\underline{V}_i^2 \leq c_{ii} \leq \overline{V}_i^2, \quad i \in \mathcal{N} \quad (3.25c)$$

$$c_{ij} = c_{ji}, \quad s_{ij} = -s_{ji}, \quad (i, j) \in \mathcal{E} \quad (3.25d)$$

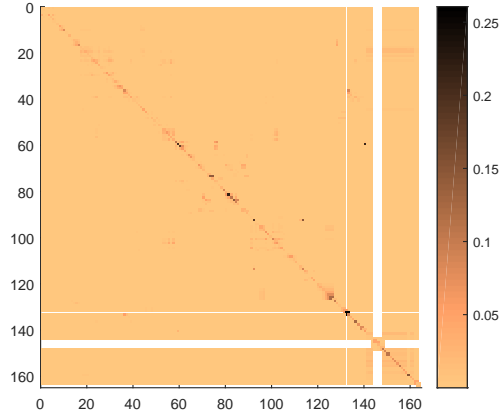
$$c_{ij}^2 + s_{ij}^2 \leq c_{ii}c_{jj} \quad (i, j) \in \mathcal{E} \quad (3.25e)$$

$$(3.8d), (3.8e), (3.22)$$

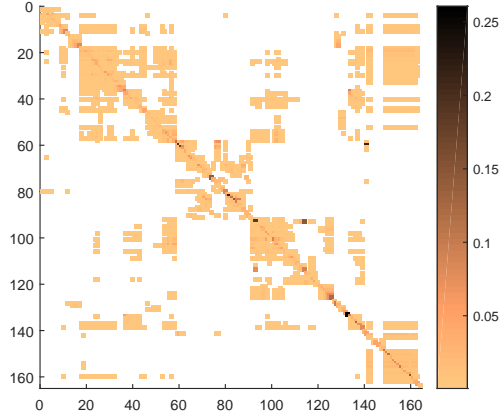
where the power flow equations (3.8b)-(3.8c) are rewritten in the c, s variables as (3.25a) and (3.25b). $N(i)$ denotes the set of buses adjacent to bus i . The line real and reactive powers are $P_{ij} = G_{ij}c_{ij} - B_{ij}s_{ij}$ and $Q_{ij} = -G_{ij}s_{ij} - B_{ij}c_{ij}$. The nonconvex constraint (3.24) is relaxed as (3.25e), which can be easily written as an SOCP constraint as $\|[c_{ij}, s_{ij}, (c_{ii} - c_{jj})/2]^T\|_2 \leq (c_{ii} + c_{jj})/2$. (3.25c) is a linear constraint in the square voltage magnitude c_{ii} . Notice that the SOCP formulation of the voltage stability constraint (3.22) is not a relaxation, but an exact formulation of the original voltage stability condition (3.11b), and it fits nicely into the overall SOCP relaxation of the VSC-OPF model (3.25). We have employed the basic SOCP relaxation of the AC-OPF in (3.25). There are many ways to strengthen the relaxation, see [85] for a few formulations. The main advantage of the adopted formulation lies in its speed, which may proven crucial for certain online applications. On the other hand, the main point we try to convey in the paper regarding the convex formulation is that the proposed voltage stability constraint is in fact second-order cone representable. This simple fact means that the constraint can be integrated in any other SOCP relaxation as well.

3.4.3 Sparse Approximation of SOCP Relaxation

Due to the density of stability condition (3.22a), the computation times of the VSC-OPF formulation (3.25) are significantly longer than normal OPF especially for larger power systems. The differences in computation time can be observed from Table I, where it is seen that for large IEEE instances, VSC-OPF is much slower than AC-OPF. The term ‘density’ refers to the fact that each voltage stability constraint in (3.22a) is coupled with almost all load buses since the matrix A in (3.22a) is dense. This is to be contrasted with the power flow equations or line flow constraints where the admittance matrices are sparse and power injection of a bus is only a function of its voltage phasor as well as those of its neighboring buses. Fig. 3.2 shows the sparsity pattern of the matrix A for IEEE 300-bus system, it can be seen from Fig. 3.2a that almost all entries are nonzero



(a) Original Sparsity Pattern.



(b) Sparsity Pattern with Entries Less Than 5×10^{-4} Set to Zero.

Figure 3.2: Sparsity Pattern of Matrix A for IEEE 300-Bus System

even though most of them are very small. To better discern the relative magnitude of the entries, we set entries less than 5×10^{-4} to zero in Fig. 3.2b, now the heat map becomes much sparser which suggests that a majority of entries are indeed small ($< 5 \times 10^{-4}$). Therefore, in order to speed up computation, we can approximate most of the entries by constants without sacrificing too much accuracy.

The first step of the approximation is to approximate the coefficient matrix A of the stability constraints by a sparse matrix \tilde{A} . To illustrate our approach of sparse approximation, we rewrite the linear constraint (3.22a) in matrix-vector form as

$$x - Ay \geq \underline{t}. \quad (3.26)$$

Then the approach to construct the sparse approximate matrix \tilde{A} can be summarized as in Algorithm 3.

Simply put, for each row of matrix A , Algorithm 3 constructs the corresponding row of the ap-

Algorithm 3 Sparse approximation of A

```

1:  $\gamma \leftarrow \gamma_0$  {initialize tunable sparsity parameter}
2:  $\tilde{A} \leftarrow \mathbf{0}_{n \times n}$  {initialize  $\tilde{A}$ }
3: for  $1 \leq i \leq n$  do
4:    $RS \leftarrow \sum_j A_{ij}$  {compute  $i$ th row sum of matrix  $A$ }
5:   while  $\sum_j \tilde{A}_{ij} < \gamma RS$  do
6:      $j_{\max} \leftarrow \arg \max_j a_{ij}$ 
7:      $\tilde{A}_{i,j_{\max}} \leftarrow A_{i,j_{\max}}$ 
8:      $A_{i,j_{\max}} \leftarrow 0$ 
9:   end while
10: end for

```

proximate matrix \tilde{A} by ignoring all elements except the largest ones whose sum amounts to more than γ of the total row sum. We notice that the element Z_{ij} of the impedance matrix can be understood as the coupling intensity measure between buses i and j . Thanks to the sparsity of practical power systems, each bus is only strongly coupled with its neighboring buses and weakly coupled with most other buses. Therefore, the matrix \tilde{A} is generally sparse. We notice a similar approximation has been applied to the L -index in the context of PMU allocation [117]. The connection between L -index and the proposed stability condition has been discussed in Section 3.3 and more extensively in [150].

Then (3.26) can be approximated by

$$x - \tilde{A}y \geq \underline{t} + \Delta a / \bar{V}, \quad (3.27)$$

where $\Delta a \in \mathbb{R}^n$ is the row sum difference between A and \tilde{A} that is defined as $\Delta a_i = \sum (a_i - \tilde{a}_i)$ and $\bar{V} = \max\{\bar{V}_i \mid i \in \mathcal{N}_L\}$. We have thus obtained the sparse VSC-OPF formulation which is identical to (3.25) except the stability constraint (3.22a) is replaced by (3.27). The new formulation is presented as

$$\begin{aligned} \min \quad & \sum_{i \in \mathcal{N}_G} f_i(P_{G_i}) \\ \text{s.t.} \quad & P_{G_i} - P_{D_i} = G_{ii}c_{ii} + \sum_{j \in \mathcal{N}(i)} P_{ij}, \quad i \in \mathcal{N} \end{aligned} \quad (3.28a)$$

$$Q_{G_i} - Q_{D_i} = -B_{ii}c_{ii} + \sum_{j \in \mathcal{N}(i)} Q_{ij}, \quad i \in \mathcal{N} \quad (3.28b)$$

$$\underline{V}_i^2 \leq c_{ii} \leq \bar{V}_i^2, \quad i \in \mathcal{N} \quad (3.28c)$$

$$c_{ij} = c_{ji}, \quad s_{ij} = -s_{ji}, \quad (i, j) \in \mathcal{E} \quad (3.28d)$$

$$c_{ij}^2 + s_{ij}^2 \leq c_{ii}c_{jj} \quad (i, j) \in \mathcal{E} \quad (3.28e)$$

$$(3.8d), (3.8e), (3.22b)–(3.22d), (3.27)$$

We notice that feasibility of problem (3.28) is implied by the feasibility of the original problem

Table 3.1: Results Summary for Standard IEEE Instances.

Test Case	Cost (\$/h)		OG (%)	\underline{t}	t_a^{AC}	$\Delta\lambda_{\max}^{AC} (%)$	$\Delta\sigma_{\min}^{AC} (%)$	DS (%)
	AC	SOCP						
case24_ieee_rts	64059.32	63344.99	1.12	0.86	0.86	0.12	0.16	0.08
case30	577.16	574.90	0.39	0.97	0.97	5.02	0.00	0.07
case_ieee30	9985.41	9220.51	7.66	0.88	0.88	7.92	3.75	0.60
case39	43667.91	42552.76	2.55	0.83	0.83	6.49	0.32	0.48
case57	41737.79	41710.91	0.06	0.66	0.66	0.02	0.02	0.31
case89pegase	5849.28	5810.12	0.67	0.72	0.72	2.22	0.21	2.61
case118	130009.61	129385.66	0.48	0.98	0.98	-0.21	0.33	0.44
case300	724935.75	718655.31	0.87	0.29	0.29	-0.30	1.13	1.03
case1354pegase	74062.27	74000.28	0.08	0.64	0.64	0.87	0.00	0.93
case2383wp	1857927.67	1846897.40	0.59	0.77	0.77	0.00	0.00	1.64
average	—	—	1.45	0.76	0.76	1.99	0.59	0.82

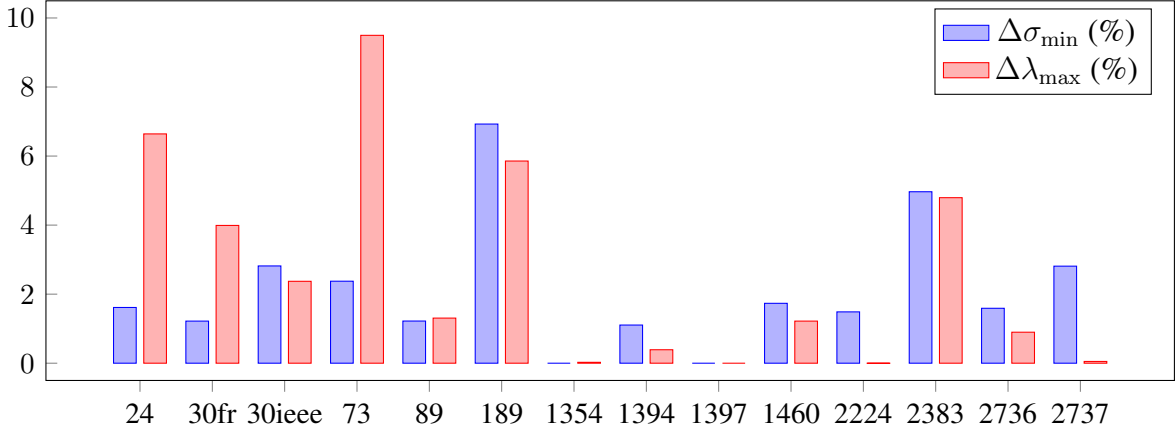


Figure 3.3: Results Summary for NESTA Instances From Congested Operating Conditions.

(3.25). To see this we only need to focus on (3.27) and (3.22a), from which we have

$$\begin{aligned}
 x - Ay &\leq x - \tilde{A}y - (\Delta a)y_{\min} \\
 &\leq x - \tilde{A}y - \Delta a/\bar{V},
 \end{aligned}$$

where the last inequality comes from (3.18), (3.19), (3.25c).

3.5 Computational Experiments

In this section, we present extensive computational results on the proposed VSC-OPF model (3.11), its SOCP relaxation (3.25), and the sparse approximation (3.28) tested on standard IEEE instances available from MATPOWER [158] and instances from the NESTA 0.6.0 archive [36]. The code is written in MATLAB. For all experiments, we used a 64-bit computer with Intel Core i7 CPU 2.60GHz processor and 4 GB RAM. We study the effectiveness of the proposed VSC-OPF on

achieving voltage stability, the tightness of the SOCP relaxation for the VSC-OPF, as well as the speed-up and accuracy of the sparse approximation.

Two different solvers are used for VSC-OPF:

- Nonlinear interior point solver IPOPT [147] is used to find local optimal solutions to VSC-OPF.
- Conic interior point solver MOSEK 7.1 [1] is used to solve the SOCP relaxation of VSC-OPF.

3.5.1 Method

Below we briefly describe the methodologies used in this section to evaluate and demonstrate the effectiveness of the proposed VSC-OPF formulation.

Evaluating the Performance of the Proposed VSC-OPF

During normal operating conditions, the voltage stability condition (3.9) is normally satisfied. That is, at least for lightly loaded IEEE test cases in MATPOWER, the constraint (3.11b) with small \underline{t} will not be binding. This is to be expected, since the stability margins of systems under normal operating conditions are relatively high. To evaluate the formulation in a more meaningful way, we set the margin threshold \underline{t} as follows.

To determine the voltage stability threshold in (3.11) for each test instance, we first solve a *minimum threshold maximization problem*. That is, we maximize the minimum value of the left hand side of (3.11b) among all load buses subject to power flow, nodal voltage, and generation constraints (3.8b) – (3.8f). The threshold \underline{t}_i in (3.11b) is set as the slightly decreased maximum threshold from the optimal objective value. In this way, we try to force the voltage stability constraint (3.11b) to be binding and examine the effect of restraining a high \underline{t}_i on system voltage stability improvement.

For comparison, we also solve a *relaxed OPF problem* for each test instance, which is the same as (3.11) except that the voltage stability constraint (3.11b) is unbounded. Two voltage stability indices, i.e. the MSV of the reduced power flow Jacobian J_{LL} and the loading margins to voltage instability of the VSC-OPF formulation (3.11) and the relaxed OPF problem are compared. It is expected that constraint (3.11b) restrains system stability level such that level of stability is improved and voltage stability indices for the VSC-OPF formulation are superior to that of the relaxed OPF problem.

Recovering Bus Voltage Phasors from SOCP Relaxation

To evaluate the SOCP relaxation (3.25), in addition to examine the optimality gap, we compare the MSV obtained by solving (3.25) with the one obtained from the original problem (3.11), which requires the recovery of nodal voltages. We know the variables c_{ii} are simply the squared bus

voltage magnitudes, so bus voltage magnitudes can be directly recovered from SOCP results. To recover voltage phase angles, we use the following relationship:

$$A_{inc}^T \theta = b \quad (3.29)$$

where A_{inc} is the bus incidence matrix and b is the vector of phase angle differences which can be calculated from SOCP results as $b_k = \text{atan2}(s_{ij}, c_{ij})^1$ if (i, j) is the k th branch in \mathcal{E} . Denote the number of buses by $n_g := n + m$ and number of branches by n_ℓ , then $n_\ell > n_g$ for almost all meshed networks, and the system (3.29) is overdetermined. We find the least squares solution of (3.29) through pseudoinverse of the bus incidence matrix:

$$\tilde{\theta} = (A_{inc}^T)^\dagger b. \quad (3.30)$$

Therefore the phase angle of bus voltages can be recovered and the voltage at bus i is given as $V_i = \sqrt{c_{ii}} e^{i\tilde{\theta}_i}$. The recovered voltages will be used to calculate the MSVs of the SOCP-VSC-OPF results.

3.5.2 Results and Discussions

The results of our computational experiments on VSC-OPF and its SOCP relaxation are presented in Table 3.1 and Fig. 3.3 for standard IEEE and NESTA instances, respectively. The “Cost” columns in Table 3.1 shows the objective values of the VSC-OPF model (3.11) and its SOCP relaxation (3.25). In addition, six sets of information are provided in Table 3.1:

- OG(%) is the percentage optimality gap between the lower bound LB of the objective value obtained from the SOCP relaxation of VSC-OPF (3.25) and an upper bound UB obtained from (3.11) by IPOPT. It is calculated as $100\% \times (1 - LB/UB)$.
- \underline{t} is the fixed voltage stability threshold used in the optimization problem (right hand side of (3.11b)).
- t_a^{AC} is the minimum value of $|V_i| - \sum_{j=1}^n A_{ij}/|V_j|$ for all load bus i calculated *after* solving VSC-OPF (3.11).
- $\Delta\lambda_{\max}^{AC}(\%)$ is the percentage increase of loading margins of VSC-OPF (3.11) (λ_1) and that of its relaxed OPF counterpart (λ_2) calculated as $100\% \times (\lambda_1/\lambda_2 - 1)$. The loading margin is the maximum loading multiplier such that the power flow Jacobian remains nonsingular under proportional load and generation increase. They are calculated using the Continuation Power Flow tool in MATPOWER.

$$^1 \text{atan2}(y, x) = \begin{cases} \arctan \frac{y}{x} & x > 0 \\ \arctan \frac{y}{x} + \pi & y \geq 0, x < 0 \\ \arctan \frac{y}{x} - \pi & y < 0, x < 0 \\ +\frac{\pi}{2} & y > 0, x = 0 \\ -\frac{\pi}{2} & y < 0, x = 0 \\ \text{undefined} & y = 0, x = 0 \end{cases}$$

- $\Delta\sigma_{\min}^{AC}(\%)$ is the percentage increase of MSV of the reduced power flow Jacobian of VSC-OPF (3.11) (σ_1) and that of its relaxed OPF counterpart (σ_2) calculated as $100\% \times (\sigma_1/\sigma_2 - 1)$.
- DS(%) is the percentage difference between the MSV σ_{\min}^{AC} obtained from AC-OPF (3.11) and the MSV σ_{\min}^{SOCP} from the SOCP relaxation. calculated as $100\% \times |\sigma_{\min}^{SOCP}/\sigma_{\min}^{AC} - 1|$.

Stability Margin Improvement

As shown by Table 3.1, on average, the proposed VSC-OPF model improves the loading margin by about 2% for the IEEE instances over the relaxed OPF problem with unbounded voltage stability constraint. We also see that several instances have significantly larger improvements. For example, case30, case.ieee30 of 30-bus system and case39 of 39-bus system all have more than 5% improved loading margins; it is seen from Fig. 3.3 that several instances in the NESTA archive have more than 5% loading margin increase as well, for instance 24-bus, 73-bus, and 189-bus systems. It is worth noting that there are two IEEE instances (118-bus and 300-bus systems) where the loading margins decrease. This does not necessarily mean the system voltage stability level is worsen as the loading margin is only measured along a specific ray of loading variation. In fact, the MSVs of the two instances both increase, suggesting the overall stability condition may be improved.

As for the MSV, we see from Table 3.1 that for IEEE instances the increase are all nonnegative, with an average value of 0.59%. This is consistent with our discussion in Section 3.3 that the voltage stability constraint (3.11b) helps preserve the diagonal dominance of the transformed power flow Jacobian. In fact, the increase of MSVs for NESTA instances are all nonnegative as well. In addition, we see from Fig. 3.3 that there are a few instances that experience large MSV increase, notably 189-bus and 2383-bus systems. We also see that there are instances for which both $\Delta\sigma_{\min}$ and $\Delta\lambda_{\max}$ are small, which may indicate that the relaxed OPF problems already yield solutions that have high voltage stability levels.

Tightness of SOCP Relaxation

Table 3.1 shows the average optimality gap between the SOCP relaxation (3.25) and a local solution of the non-convex VSC-OPF (3.11) is about 1.45%. The optimality gap is quite small, but still larger compared with the standard OPF. This can be attributed to the fact that the flow limits for IEEE instances are high and most of them are not binding in standard OPF, while the voltage stability constraints for VSC-OPF are binding in our experiment.

Effect of Sparse Approximation

The result summary of our computational experiments on the sparse approximation of VSC-OPF for large NESTA instances are presented in Table 3.2. The sparsity parameter in Algorithm 3 is chosen to be 0.98. The “Time” columns in the table show the computation time of the VSC-OPF model (3.25) and the sparse approximation (3.28). In addition, the table reports two sets of data as described below:

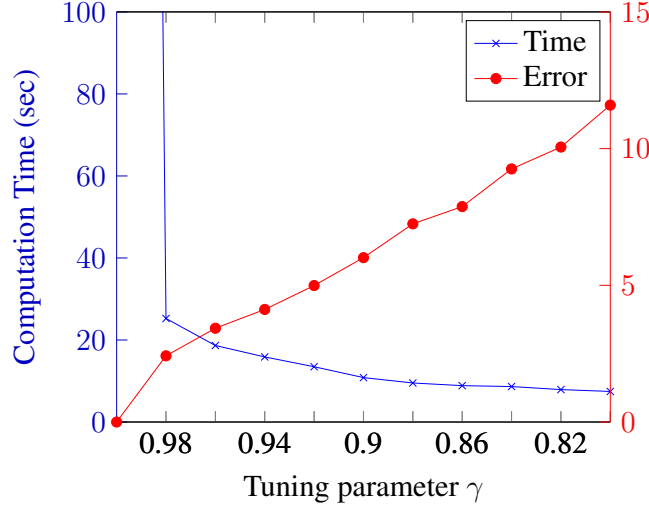


Figure 3.4: Sparse Approximation of NESTA 2737-Bus Test System

- DCT(%) is the percentage time difference between the computation time ct_n of (3.25) and ct_s of (3.28). It is calculated as $100\% \times (1 - ct_s/ct_n)$.
- DC(%) is the percentage difference between the objective value c^{SOCP} of the model (3.25) and c^s of (3.28) calculated as $100\% \times |c^s/c^{SOCP} - 1|$.

For NESTA systems with less than five buses, the sparse approximation (3.28) and the original SOCP relaxation model (3.25) are exactly the same. For system sizes ranging between 6 buses and 300 buses, the computation times of model (3.25) are sufficiently short (less than 2 seconds), which render the sparse approximation unnecessary. However, for systems with more than 1000 buses, the sparse approximation brings about significant speed-up. In fact, the speed-ups are above 90% for all instances with more than 2000 buses and the optimal solutions are obtained in less than 30 seconds for all instances. Our simulation experiments suggest that the solution accuracies are extremely high. For larger systems with more than 1000 buses, the differences of cost between (3.25) and (3.28) are all less than 0.01%.

Fig. 3.4 presents the results of our computational experiments on the sparse approximation of VSC-OPF for NESTA 2737-bus test instance with varying γ . This specific test instance is chosen since it is the largest instance we have experimented with and also the one that takes the longest computation time. In fact, it takes almost 1200 seconds to compute the optimal solution for the test instance. In the figure, blue cross shows the computation time and red dot shows the relative error of the MSV results. The relative error is calculated as $|\sigma_1 - \sigma_\gamma|/|\sigma_1 - \sigma_0|$ where σ_1 , σ_0 and σ_γ are the MSVs given by SOCP relaxation (3.25) ($\gamma = 1$), relaxed OPF problem ($\gamma = 0$), and sparse approximation with tuning parameter γ . For this test instance, $\sigma_1 \approx 0.451$ and $\sigma_0 \approx 0.439$, it can be seen from Fig. 3.3 that there is an approximately 3% increase from σ_0 to σ_1 . We see from Fig. 3.4 that the computation time sees a drastic decrease with a very small deviation of the tuning parameter from 1. Even with γ as high as 0.98, the computation time can be reduced to within 30 seconds, while further decrease in γ reduces the computation time but to a lesser extent, and the computation time gradually stabilizes at around 10 seconds. The relative error increases almost

Table 3.2: Results Summary of Sparse Approximation for Large NESTA Instances From Congested Operating Conditions.

Test Case	Time (sec)		DCT (%)	DC (%)
	Normal	Sparse		
nesta_case1354_pegase	25.04	4.24	83.05	0.00
nesta_case1394sop_eir	39.44	9.75	75.27	0.00
nesta_case1397sp_eir	40.42	10.34	74.42	0.00
nesta_case1460wp_eir	39.95	10.54	73.61	0.00
nesta_case2224 edin	274.68	22.23	91.91	0.00
nesta_case2383wp_mp	90.54	6.92	92.35	0.00
nesta_case2736sp_mp	496.59	14.92	97.00	0.00
nesta_case2737sop_mp	1190.98	25.24	97.04	0.00
average	274.71	13.02	85.58	0.00

linearly with the decrease of γ . For $\gamma = 0.98$, the relative error is only around 3%.

3.5.3 Comparison with Alternative VSC-OPF Formulation

In this section, we compare the proposed VSC-OPF formulation with an alternative formulation proposed in [156]. The VSC-OPF employed in [156] is based on Voltage Collapse Proximity Index (VCPI), a voltage stability index quantifying power transfer margins of individual branches. The VCPI index for a branch is defined as

$$\text{VCPI} = \frac{P_r}{P_{r,\max}}, \quad (3.31)$$

where P_r is the real power transferred to the receiving end, $P_{r,\max}$ is the maximum real power that can be transferred to the receiving end assuming the voltage at the sending end is fixed. It is known from the definition that $0 \leq \text{VCPI} \leq 1$ and high VCPI signifies a system that is more stressed. Let the sending and receiving end bus voltages be $|V_s|e^{i\theta_s}$ and $|V_r|e^{i\theta_r}$, and let $V_d := |V_s|e^{i\theta_s} - |V_r|e^{i\theta_r}$, then the index can be represented by the two voltages as

$$\text{VCPI} = \frac{2|V_r||V_d|}{|V_s|^2} + 2\frac{|V_r|\cos(\theta_s - \theta_r)}{|V_s|} - 2\frac{|V_r|^2}{|V_s|^2} \quad (3.32)$$

The resulting VSC-OPF formulation is the same as (3.11) except that the constraint (3.11b) is replaced with $\text{VCPI}_{\max} \leq \text{VCPI}_{\text{limit}}$ where VCPI_{\max} is the maximum VCPI among all branches and $\text{VCPI}_{\text{limit}}$ is a preset threshold. We would like to point out that the VCPI index is heuristic in nature since it has been shown in [62] that maximum branch flows are generally encountered well before the onset of voltage instability.

The results of (3.11) as well as that of the above formulation based on VCPI depend on the preset threshold. It is difficult to choose comparable thresholds for the two indices that represent similar system stress levels, since after all the effect of the indices in reflecting system stress level is what we want to investigate. Therefore, we propose to compare the two indices by formulating the

Table 3.3: Comparison of the Effect of Voltage Stability Improvement of Different VSC-OPF Formulations.

Test Case	$\Delta\lambda_{\max} (\%)$		$\Delta\sigma_{\min} (\%)$	
	(P_C)	(P_{VCPI})	(P_C)	(P_{VCPI})
case24_ieee_rts	9.87	5.21	0.32	-0.32
case30	15.06	-0.59	0.01	-5.13
case_ieee30	9.02	6.02	4.05	3.38
case39	8.96	1.39	0.51	-4.01
case57	0.52	-1.66	0.52	-0.76
case89pegase	1.39	0.99	1.27	-5.99
case118	38.43	27.57	1.58	-4.70
case300	3.33	1.63	3.43	-2.37
case1354pegase	0.92	-4.08	0.05	-4.69
case2383	2.64	— [†]	0.80	-1.63
average	9.01	4.05[‡]	1.25	-2.62

[†]Power flow based on optimization solution does not converge

[‡]Average over the first nine cases

‘voltage stability improvement’ problem in [156]. That is, instead of using the voltage stability index as a constraint, we directly optimize the sum of stability indices, subject to power flow equations, nodal voltage and power generation constraints (3.8b) – (3.8f). We then denote the two optimization problems as (P_{VCPI}) and (P_C) , since they optimize the sum of VCPI and C-index ([150]), respectively.

One thing we notice with (P_{VCPI}) is that for almost all instances, the problems experience very slow convergence: they do not converge after 1,000 iterations in IPOPT. This is probably due to the poor numerical properties of the VCPI index (3.32), since the gradient and hessian of the constraint involve the reciprocal of V_d , which is almost zero when V_s and V_r are close. We end up with the MATLAB function `fmincon` with interior-point solver as was used in [156] for (P_{VCPI}) with the default maximum number of iterations of 3,000. The program terminates with a feasible solution that is best possible, to which we compare with (P_C) solved with IPOPT. The results are shown in Table 3.3. It is seen that the proposed approach outperforms the one in [156].

3.6 Conclusions

We have presented a sufficient condition for power flow Jacobian nonsingularity and shown that the condition characterizes a set of voltage stable solutions. A new VSC-OPF model has been proposed based on the sufficient condition. By using the fact that the load powers are constant in an OPF problem, we reformulate the voltage stability condition to a set of second-order conic constraints in a transformed variable space. Furthermore, in the new variable space, we have formulated an SOCP relaxation of the VSC-OPF problem as well as its sparse approximation. Simulation results show that the proposed VSC-OPF and its SOCP relaxation can effectively restrain the stability stress of the system; the optimality gap of the SOCP relaxation is slightly larger than that of the standard OPF problem on IEEE instances due to tightness of the constraints; and the sparse

approximation yields significant speed-up on larger instances with small accuracy compromise. It has also been shown that the proposed method outperforms existing one in terms of effectiveness and computational properties.

4. Optimal Rate Design in Modern Electricity Sectors

4.1 Introduction

Electricity sectors are experiencing two major shifts: accelerating deployment of *intermittent renewable generation*¹ and pervasive information and communications technologies (ICT) [71]. The management of the demand, or *demand response*, is broadly seen as an important resource in the presence of intermittent renewables, and ICT has a clear role to play in providing information or control signals to end-use devices to enable this resource [77]. There are two types of demand response programs. One includes programs in which a utility can directly alter the demand level of a customer.² The other encompasses programs where price signals are the means the utility uses to influence its demand [56]. In this paper we focus on the latter type of programs, commonly known as rates or tariffs.

In many jurisdictions, a regulated utility distributes the electricity to most end customers. In these instances, this agent collects its revenue through a set of rates or tariffs, which are under the oversight of the regulatory body. A central element defining these tariffs is their structure or design, which specifies what charges compose these instruments [76]. Given the new reality of the electricity industry, where renewable generation and distributed energy resources (DERs) play an increasingly important role, regulators are exploring more sophisticated tariff structures [134]. Advanced rate designs that align better with marginal costs (e.g., time-varying rates) can reduce generation fuel costs and decrease investment in distribution, transmission and generation infrastructure [77]. However, when other public goals—such as carbon emission reductions—are at stake and the presence of intermittent renewables is significant, the case for advanced rates is less clear-cut. Depending on the characteristic of the system, these rate designs may or may not decrease emissions [67]; simpler tariffs structures, such as a flat rate (FR), may produce greater environmental benefits and even improve consumer surplus [87].

The emergence of distributed energy resources further complicates the analysis of rate structures. Advanced metering infrastructure (AMI), a set of technologies allowing utilities to collect and transmit granular consumption information, and home energy management systems (HEMS) enable the implementation of sophisticated rate designs, and can boost consumer price-responsiveness [53]. However, these DERs are not necessarily cost-effective. While AMI comes along with operational savings, including a decrease in meter reading or fault detection expenditures, these are insufficient to cover its capital costs ([52], [53]). Improved tariff designs in combination with HEMS could fill this gap. But they are not exempt of costs either. Implementing complex rate structures requires at least creating awareness in the population and educating retail customers on the benefits these rates; home energy management systems, that can enlarge the benefits of advanced rates, such as smart thermostats or in-home displays (see [55]), require meaningful capital outlays as well.

Given these benefits and costs, to what extent it is beneficial for a system to implement more sophisticated rate structures? While this policy-relevant question has been traditionally approached by economists, we believe that the OR community has an important role to play in improving

¹Intermittent renewables include wind and solar power plants.

²In the US, an example of this type of program is *direct load control*. Utilities can remotely control some of the devices of a customer under this program.

past answers, specially in the face of the complexities unfolding in the electricity industry. [133] gives a step in this direction. The paper investigates the impacts of different tariff structures on the efficient operation of Plug-In Hybrid Electric Vehicles (PHEVs). Using a Unit Commitment and Vehicle-Charging models, the paper finds that a simple flat rate can outperform time-varying rate designs. [133] detailed model illuminates the inefficiencies that can emerge at the retail level when using time-varying structures in sectors with non-convex costs—where production technologies have costs with terms independent of the level of production [107]. This finding contradicts economists common wisdom, which asserts that rate designs that reflect marginal costs better are more desirable (see, e.g., [77]).

[87] is another example showing that capturing new complexities, characteristic of modern electricity sectors, can challenge intuition. The paper investigates the interaction between two rate designs, a flat and a time-varying structure, and investment in the presence of renewable generation. Using a model that captures the intermittency of renewables and the pricing behavior of the utility, the paper finds that the flat rate design leaves consumers always better-off relative to the time-varying structure. Although this conclusion rests upon the characteristics of the setting that [87] analyze, the result does provide additional evidence of the importance of detailed modeling when evaluating tariff structures.

The present work adds to the contributions by developing a general technique to evaluate rate structures. We take as a starting point empirical methods which have been used to conduct welfare analysis of rate changes (e.g., [2], [30], [60], [69], [91], [111], [12], [15], [139] and [4]). We add layers of detail that capture salient aspects of modern electricity sectors. Specifically, we extend previous techniques in four significant ways. First, our method improves the consistency of the comparison among rates. Existing approaches compare tariff structures either specifying *ex-ante rate levels* (the specific values of each of the charges composing the rate structure), or imposing unnecessary constraints on them (e.g., [2], [12], [133]). In contrast, our technique computes rate levels following an optimality criterion. Researchers compare the best case of a rate structure against the best case of another (see Subsection 4.3.1 for more details).

In addition, we improve the supply cost representation of previous methods. With the exception of [133], past work either represents these costs with stylized models (e.g., [12], [87]) or omits the supply side, only assessing welfare impacts on end customers (e.g., [91], [4]). The present technique permits researchers to include complex representations of the supply side. The main difference with the approach of [133] is that we consider long term costs while this author focuses on the short-run.

A third extension allows comparing multiple rate structures simultaneously. Except for [12] and [15],

The last element that distinguish our method from others is the endogenous computation of an optimal *demand mix*. That is, the model we use for comparing rates finds the socially optimal fraction of customers enrolled in different programs. [15] explores the welfare implications of alternative demand mixes. Specifically, the paper studies the impacts of exogenously varying the fraction of customers under an advanced tariff. It shows that while the marginal benefit of increasing this fraction decreases, the marginal cost remains constant. Consequently, the authors observe that the optimal fraction ultimately depends upon the specific characteristics of the system under

study. The endogenous computation of an optimal demand mix simplifies the analysis in [15]. More importantly, in combination with the other improvements we introduce, it allows comparing portfolios of rate structures making far less assumptions than one would have to if using other approaches.

Our technique uses a nonlinear optimization model that we build based on the *Peak-Load Pricing* theory. This strand of utility pricing was developed in the seventies and eighties with the work of [135], [9], [46], [39], [109], [75], [25] and [31], and more recently was revisited by [159] and [32]. While it originally intended to provide theoretical guidelines for the optimal pricing of public utility services [38], other authors have used this theory as a framework to analyze a range of regulatory issues. [73] uses Peak-Load Pricing to construct a benchmark to understand the implications of competition at the retail level in the electricity industry; [159] explores the incentives to invest of strategic firms participating in markets where demand is fluctuating and storage is prohibitively expensive; and [32] uses the theory to explore the interaction between time-varying rates and intermittent renewables. Following the approach of these papers, we use Peak-Load Pricing as a basis for our model which we modify to meet our purposes.

4.2 Peak-Load Pricing: An Overview

The problem that this theory addresses is how to price the set of commodities that a regulated monopoly provides. It answers this question taking the perspective of a regulator. Optimal prices are such that the societal welfare is maximized subject to the revenue sufficiency and technical constraints of the regulated monopoly [38]. We formalize the Peak-Load Pricing problem following [38] and [73].

Before starting, we introduce some notation. Let Ω be a discrete sample space, q_ω the probability that $\omega \in \Omega$ occurs, and $E[\cdot]$ the associated expectation operator. We refer to an element in Ω as outcome or state of nature, and distinguish a random from a deterministic variable placing a bar on top of the former. Given a random variable \bar{y} , we denote y_ω the realization of this variable when ω occurs. The symbol \top indicates the transpose of a vector.

The theory considers a monopolist offering a set $\{1, \dots, T\}$ of goods. Customers are of different types $i \in I$, and distribute according to a frequency function δ_i , denoting the number of types i . A quasi-linear utility $U_\omega^i(d) + m_\omega$ characterizes the preferences of the customers of type i over consumption bundles $d \in \mathbb{R}_+^T$. The scalar m_ω is her expenditure in all other goods. For a vector of prices $p_\omega \in \mathbb{R}_+^T$, a customer with an income M_i consumes according to the demand function $D_\omega^i(p_\omega) := \arg \max_{d \geq 0} \{U_\omega^i(d) + M_i - p_\omega^\top d\}$.³ It is customary to assume that $U_\omega^i(\cdot)$ is strictly concave and, thus, $D_\omega^i(p)$ is a singleton. The *gross surplus* of this customer is $S_\omega^i(p_\omega) := U_\omega^i(D_\omega^i(p_\omega))$.

We define $\bar{D}^{I'}(\bar{p})$ as the aggregated consumption of types $i \in I'$, where $I' \subseteq I$. That is, $\bar{D}^{I'}(\bar{p}) = \int_{I'} \bar{D}^i(\bar{p}) \delta_i di$. Defined similarly, $S^{I'}(p_\omega)$ represents the aggregated gross surplus.

The monopolist offers a two-part rate structure. That is, a contract which has a fixed charge l and a vector of volumetric charges \bar{p} . In this arrangement, the monopolist charges $p_{t\omega}$ per unit of

³The problem the customer solves is $\{U_\omega^i(d) + m_\omega : M_i \geq p_\omega^\top d + m_\omega\}$. Because peak-load pricing assumes $m_\omega > 0$ in any optimum, one can simplify the problem.

consumption of good t under ω . The corresponding consumer surplus is

$$CS(l, \bar{p}) = E [\bar{S}^I(\bar{p}) - \bar{p}^\top \bar{D}^I(\bar{p})] - l \cdot \nu_I, \quad (4.1)$$

where $\nu_{I'} := \int_{I'} \delta_i di$ for any $I' \subseteq I$.

The monopolist produces with a set of technologies that we index with the letter $k \in K$. Each technology differs from others on its variable costs per unit of production, $c_{\omega k} \in \mathbb{R}_+^T$, its fixed costs \hat{r}_k , and its availability factor, $\rho_{\omega k} \in \mathbb{R}_+^T$, capturing the variability in the technology's availability—e.g., due to the intermittent output of some renewables or the occurrence of outages. Before uncertainty realizes, the monopolist decides the installed capacity of each technology, x_k . After, the firm determines a production vector for each technology, $y_{\omega k} \in \mathbb{R}_+^T$. For a consumption vector \bar{d} , the monopolist's cost function satisfies

$$C(\bar{d}) = \min_{(x, y)} \sum_{k \in K} E [\bar{y}_k^\top \bar{c}_k + x_k \hat{r}_k] \quad (4.2)$$

subject to

$$\bar{d} \leq \sum_{k \in K} \bar{y}_k, \quad (4.3)$$

$$0 \leq \bar{y}_k \leq x_k \bar{\rho}_k, \quad k \in K \quad (4.4)$$

and the firm's profit is

$$\Pi(l, \bar{p}) = E [\bar{p}^\top \bar{D}^I(\bar{p})] + l \cdot \nu_I - C(\bar{D}^I(\bar{p})) - \Pi_0, \quad (4.5)$$

where Π_0 captures transmission and distribution costs, overhead expenses and the opportunity cost of the monopolist.

The welfare maximization problem or, as it is referred to in Peak-Load Pricing (e.g., [38], [73]), the *Ramsey* problem is

$$\max_{(l, \bar{p})} \{CS(l, \bar{p}) : \Pi(l, \bar{p}) \geq 0, (l, \bar{p}) \in \mathcal{L} \times \mathcal{P}\}. \quad (4.6)$$

Henceforth, we refer to $\mathcal{L} \times \mathcal{P}$ as rate structure, and to an element of this set as rate level.

The literature theoretically explores optimal pricing rules for alternative structures. For instance, [73] consider the case where $\mathcal{L} = \mathbb{R}$, and compare a real-time pricing structure (RTP), in which $\mathcal{P} = \mathbb{R}_+^{T \cdot |\Omega|}$, with a flat rate structure, where $\mathcal{P} = \{\bar{p} \in \mathbb{R}_+^{T \cdot |\Omega|} : p_{t\omega} = p_{t'\omega'} \forall (t, \omega)\}$; [38], on the other hand, focus on the case where $\mathcal{L} = \{0\}$, and also review the real-time and flat rate cases.

4.3 A Method to Compare Rate Structures

We propose using the model of Peak-Load Pricing as a framework to compare rate structures. For a set of tariffs under analysis, the comparison requires solving the Ramsey problem for each of the corresponding structures, and then comparing the optimal values of the problem.

This method is closely related to the approaches used by [2], [30], [60], [69], [91], [111], [139] and [4]. For a change in rates, these studies compute a change in welfare (W) using that $\Delta W = \Delta \Pi - \Delta r + \Delta Y$, where $\Delta \Pi$ is the change in producer surplus, Δr the variation in demand side infrastructure costs, and ΔY is the compensating variation—the money that when taken away from an individual leaves him with the same level of welfare he had before the price change [98]. Researchers can use the optimal value of the Ramsey problem to quantify ΔW . With the following Lemma, we formalize this claim.

Lemma 1. *Suppose prices change from (l_1, \bar{p}^1) to (l_2, \bar{p}^2) , and let v_1 and v_2 be the optimal value of the Ramsey problem for the rates 1 and 2, respectively. Then,*

$$\Delta W = v_2 - v_1 - \Delta r. \quad (4.7)$$

4.3.1 Consistently Comparing Rate Structures

One advantage of using (4.7) is that it offers a consistent criterion to compare rate structures. The solution of the Ramsey problem corresponds to the optimal rate levels. Thus, researchers compute the change in welfare associated to the best case of each structure.

This criterion offers an alternative to previous approaches. [4], [69], [91], [111] and [139] compute the welfare changes using predefined rates. [2] compare two time-of-use (TOU)⁴ with a flat rate structure. The authors determine the TOU assuming a difference between the peak and off-peak charges and imposing revenue neutrality. [30] follows a similar approach except that the authors search for an optimal TOU rate evaluating various peak to off-peak ratios. The method that is closest to ours is the one that [60] use. This work computes welfare changes from a flat rate to a TOU and to an RTP structures. As in our method, the paper numerically finds optimal rate levels for each structure. The key difference is that the authors use a simplified representation of the production costs. While a simplification, the function do captures an important trade-off present in electricity industries: a more capital intensive production mix, commonly associated to an increased average cost of capital, will tend to have lower short-run marginal costs at all levels of production.

4.3.2 A Flexible Cost Function

The difficulty of using the approach of [60] is that the cost function cannot be easily customized. It is not possible to use this function, for instance, in systems with potentially high penetrations of intermittent renewables. A crucial determinant of the value of these technologies is how their output correlates with consumption. This aspect is missing in the cost representation of [60]. Using the cost function of Peak-Load Pricing, on the other hand, avoids this problem, without missing the trade-off between capital and short-run marginal costs.

Furthermore, a researcher can easily modify (4.2)-(4.4) to increase the realism and suitability of this cost representation depending on the data available. Indeed, the technical results of Subsections 4.3.5 and 4.3.6 hold if the objective function is convex, and (4.4) is a general convex set. In

⁴A TOU structure charges differently depending on the hour of the day, day of the week and possibly season.

particular this allows modeling a transmission system, rationing and disruption costs,⁵ and technologies with storage and ramping constraints.

4.3.3 Comparing Portfolios of Rate Structures

An element characteristic of previous tariff analyses is the pairwise comparison of rate structures. With the exception of [12] and [15], past work assesses welfare changes resulting from the whole population being in one rate and then switching to another.

In practice, a utility recovers its costs offering a portfolio of tariffs. Each rate in the portfolio applies to an specific class—a fraction of the customer base with particular cost characteristics [120]. It is not uncommon for utilities to distinguish various classes (e.g., industrial, commercial and residential customers) and, thus, offer portfolios with several rates [120]. It seems then appropriate to have methods that allow measuring welfare changes when changing more than one rate at a time.

These approaches can also help improving analyses focusing on just one rate structure. Changing the tariff of one class can impact other classes as well. For instance, if a new rate reduces the contribution of the class to the aggregated peak consumption, this will cause a reduction in the overall production costs.⁶ If this effect is systematic, in the medium to long term, it will translate into lower bills for all customers, not just those in the class with the new rate.

To the best of our knowledge, only [12] and [15] explore welfare changes while simultaneously adjusting multiple rates. These papers study a setting with two type of customers, those enrolling in a flat rate structure and those in an RTP tariff. The studies analyze various scenarios in terms of the fraction of the population under each rate. For each scenario they compute rate levels that satisfy market equilibrium conditions, and calculate the corresponding welfare metric. Our method builds upon the idea of comparing multiple rates which are adjusted simultaneously to new systems' conditions. We expand the approach of [15] with a model that allows comparing general portfolios of rate structures.

In order to achieve this goal, we introduce a slight modification to the Ramsey problem. Let $\mathcal{L} := \mathcal{L}_1 \times \dots \times \mathcal{L}_n$, $\mathcal{P} := \mathcal{P}_1 \times \dots \times \mathcal{P}_n$ and $\mathcal{L} \times \mathcal{P}$ be the constraint set for the rate levels, that we call *portfolio of rate structures*. We consider \mathcal{P} being a convex set. Besides, we let $l = (l_1, \dots, l_n)^\top$ be a vector of fixed charges and $\bar{p} = ((\bar{p}^1)^\top, \dots, (\bar{p}^n)^\top)^\top$ a block vector of volumetric charges. Now the duple (l, \bar{p}) corresponds to a portfolio of rate levels. The subindex h identifies a particular rate and the partition of population under this tariff (for instance, a class). The set I_h contains the types under partition h , and α_h is the number of customer under this partition. We focus on the case where $I_h \cap I_{h'} = \emptyset$, $\forall h \neq h'$. This is the relevant case since classes distinguish customers with different characteristics. We modify the consumer surplus function of Peak-Load Pricing as follows

$$CS(l, \bar{p}) = \sum_{h=1}^n E [\bar{S}^h(\bar{p}^h) - \bar{D}^h(\bar{p}^h)^\top \bar{p}^h] - l^\top \alpha, \quad (4.8)$$

⁵Following the technique of [39].

⁶In the electricity sector more efficient units have priority. This implies that during peak periods the more inefficient plants are used, which increases the marginal cost of production.

and similarly update the profit function

$$\Pi(l, \bar{p}) = \sum_{h=1}^n E [\bar{D}^h(\bar{p}^h)^\top \bar{p}^h] + (l + r)^\top \alpha - C \left(\sum_{h=1}^n \bar{D}^h(\bar{p}^h) \right) - \Pi_0, \quad (4.9)$$

where, for simplicity, we replaced $\bar{S}^{I_h}(\bar{p}^h)$ and $\bar{D}^{I_h}(\bar{p}^h)$ by $\bar{S}^h(\bar{p}^h)$ and $\bar{D}^h(\bar{p}^h)$, respectively. In (4.9) the vector $r = (r_1, \dots, r_n)^\top$ has in its components fixed costs which are directly associated with each rate. For example, if a utility decides to implement a time-varying rate for customers in h , it needs to upgrade its metering equipment. The parameter r_h is the cost of the new capital, and may also account for program implementation and marketing costs (all expressed as annuities).

Observe that (4.8) and (4.9) do not alter the structure of the Ramsey problem, only increase its dimensionality. If it is possible to solve (4.6) efficiently, then the same applies for the new problem.

4.3.4 Optimal Demand Mix

Distributed energy resources change the way in which customers interact with the electricity grid. Households and businesses become more responsive to complex price signals [55], and they may even sell energy and provide reliability services to the grid [72]. In an effort to adapt to this emerging reality, regulators are rethinking the design of existing rates [134]. The challenge involves developing tariffs that (i) provide the right long term incentives, so that DERs are adopted efficiently [105],⁷ and (ii) uncover the operational value of these resources [90]. To enable researchers to explore the extent in which a given rate structure meets these goals, we introduce a final modification to the Peak-Load Pricing framework.

This modification builds upon the observations of [15] and [73]. These papers analyze the value of advanced metering infrastructure (AMI) as enabler of real-time pricing. They observe that, depending on the capital cost of this distributed energy resource, it is optimal to deploy AMI in only a fraction of the customer base. The reason is that the marginal benefit of an increasing number of customers enrolled in RTP decreases [15], while the marginal cost (the capital cost of AMI) remains constant. This result suggests treating the number of customers with the same rate structure and DER in a similar fashion as one treats the installed capacity of a production technology. Making an analogy with the supply side, one can think the distribution of customers across tariffs and DERs as a demand mix. The final modification we introduce to the Peak-Load Pricing model, allows comparing tariff structures under optimal supply and demand mixes.

Beyond improving the internal consistency of the rate evaluation method we propose, this symmetrical treatment improves the accuracy of the technique. Comparing rates assuming arbitrary long-run configurations for the demand can produce misleading results. Past rate analyses, such as [30] or [60], concluded that time-of-use rates were not cost-effective for residential customers. Metering costs would have outweighed efficiency gains if the program would have been implemented for the whole customer base. The observations of [15] and [73] weaken the conclusions of these studies. Time-varying rates could have passed the cost-benefit test if the authors would have

⁷In this context, the word adoption refers to households and businesses acquiring a resource relevant for the grid operation, for instance, a solar photovoltaic panel or an electric vehicle. Efficient adoption refers to the deployment of these resources at the right place and time.

focused on the optimal subset of the population. In addition, computing a demand mix is important for a reason not directly related with rate analysis. An optimal supply and demand mix provides regulators and policymakers with a snapshot of the long term configuration of the system, given a portfolio of rate structures. This perspective can be used as a benchmark and also to set targets for DERs adoption or rate enrollment.

To model a demand mix, we now assume that I is a discrete set, with ν_i being the number of customers under type $i \in I$ (e.g., $I = \{\text{industrial, commercial, residential}\}$). For customers that adopt a technology h the rate that applies is (l_h, \bar{p}^h) ; and α_h continues to be the number of customers in h . Defining the matrix Γ such that $[\Gamma]_{ih} = 1$ if a type i can enroll in h and 0 otherwise, the feasible region for α , henceforth the demand mix, is $\mathcal{A} := \{\alpha \in \mathbf{R}_+^n : \Gamma\alpha \leq \nu\}$.

As in the setting of [15], we focus on the case where $\sum_i [\Gamma]_{ih} = 1$. That is, we consider that one rate applies to only one customer type. In our setting the type i is a representative customer. Though interesting, we postpone the development of the general case, in which different customer types can enroll in the same rate, for future work.

The new consumer surplus function is now

$$CS(l, \bar{p}) = \sum_{h=1}^n \alpha_h E [\bar{S}^h(\bar{p}^h) - \bar{D}^h(\bar{p}^h)^\top \bar{p}^h] - l^\top \alpha, \quad (4.10)$$

and the utility surplus

$$\Pi(l, \bar{p}) = \sum_{h=1}^n \alpha_h E [\bar{D}^h(\bar{p}^h)^\top \bar{p}^h] + (l + r)^\top \alpha - C \left(\sum_{h=1}^n \alpha_h \bar{D}^h(\bar{p}^h) \right), \quad (4.11)$$

where r_h not only includes the costs associated to the implementation of the rate but also those related to the technology of the customers enrolled in h . For simplicity, henceforth we consider that r_h includes the costs $\Pi_0 / \sum_i \nu_i$.⁸

We simplify the Ramsey problem noting that $\Pi(l, \bar{p}) = 0$ in the optimum. Using this condition, and equations (4.10) and (4.11), the version of the Ramsey problem we propose for rate analysis is

$$\max_{(\alpha, \bar{p})} \sum_{h=1}^n \alpha_h E [\bar{S}^h(\bar{p}^h) - r_h] - C \left(\sum_{h=1}^n \alpha_h \bar{D}^h(\bar{p}^h) \right) \quad (4.12)$$

subject to

$$\bar{p} \in \mathcal{P}, \quad (4.13)$$

$$\alpha \in \mathcal{A}. \quad (4.14)$$

Researchers evaluating portfolios of rate structures can solve (4.12)-(4.14) for alternative definitions of \mathcal{P} and compare the optimal value of this problem. Note that it is no longer necessary to add exogenously the variation in demand side capital costs (Δr) to the variation in optimal values

⁸In practice utilities and customers share technology costs, for instance, utilities may own metering infrastructure and customers rooftop solar panels. However, ownership is not relevant from a social planning perspective when the focus is efficiency. Besides, utilities can always pass along this cost with the fixed charge l_h .

of the Ramsey problem, as in (4.7). These costs are part of the objective of (4.12)–(4.14).

4.3.5 Enhancing the Applicability of the Framework

The Ramsey problem is nonlinear, and non-convex. This poses two important challenges to analysts using this model. First, in general it is not possible to guarantee that a solution of (4.12)–(4.14) is globally optimal. Thus, despite finding solutions for two competing portfolios of rate structures their comparison could be inconsistent. The analyst could benchmark a sub-optimal against an optimal solution. Second, the non-convexity of the problem limits its scalability as the problem size greatly decreases the performance of non-convex solvers. To enhance the practical applicability of the method, we make the following assumption:

Assumption 2. *The gross surplus function $S_\omega^h(\cdot)$ is strictly concave and the demand function $D_\omega^h(\cdot)$ is convex.*⁹

Under Assumption 2 (4.12)–(4.14) remains nonlinear and not necessarily convex or concave. However, its specific structure allows us to develop an efficient solution method. The following proposition suggests a suitable approach.

Proposition 4. *Let P_α refer to the problem (4.12)–(4.14) with α entering as a fixed parameter, and let $g(\alpha)$ be the optimal value of P_α . Under Assumption 2, for any $\alpha \in \mathcal{A}$, $g(\alpha)$ is concave.*

Indeed, even though $g(\cdot)$ does not have an explicit functional form, we can leverage an iterative procedure solving

$$\max \{g(\alpha) : \alpha \in \mathcal{A}\} \quad (4.15)$$

to find a solution to the Ramsey problem. Because (4.15) is a convex problem, such an approach could potentially outperform non-convex solvers. In Section 4.4 we present evidence suggesting that this is in fact the case.

4.3.6 Optimal Long Term Incentives

An implicit assumption of the Ramsey problem is that regulators can control the demand mix α , i.e., the adoption of technologies and rate enrollment. While these authorities could have certain influence on the roll-out of distribution equipment, for many demand side technologies and rates customers are the agents making the adoption and enrollment decisions. At best, regulators could design long term incentives consistent with a desired demand configuration. Thus, the solution of (4.12)–(4.14) should be interpreted as the demand mix given optimal long term incentives.

For concreteness, we now propose one approach for setting these incentives. Let $\bar{p}(\alpha)$ be the optimal solution of P_α . Define, in addition, $\bar{\lambda}(\alpha)$ as the dual variable of (4.3) when the demand parameter is the aggregated demand (the argument of $C(\cdot)$ in (4.11)) evaluated at $(\alpha, \bar{p}(\alpha))$. Further, define the vector of fixed charges $l(\alpha)$ as follows,

$$l_h(\alpha) = E \left[(\bar{\lambda}(\alpha) - \bar{p}^h(\alpha))^\top \bar{D}^h(\bar{p}^h(\alpha)) \right] + r_h. \quad (4.16)$$

⁹Convex demand functions are mappings whose components are all convex.

The incentive rule we propose is *set the rate levels at $(l(\alpha), \bar{p}(\alpha))$ when the demand mix is α .*

This rule has appealing theoretical properties. It replicates the incentive structure of an industry with a distribution utility and a competitive wholesale market (see e.g., [73]). More importantly, one can show that these incentives align customers' individual choices with the maximization of societal welfare. The three results that follow formalize this property.

First, we observe that if the system is at a socially optimal configuration, under the rule we propose no customer has incentives to switch.

Proposition 5. *Let $(\alpha^*, \bar{p}(\alpha^*))$ be an optimal solution of (4.12)–(4.14), and $l(\alpha^*)$ the vector of fixed charges. Then, the portfolio of rate levels $(l(\alpha^*), \bar{p}(\alpha^*))$ is such that no customer has an incentive to switch to an alternative rate.*

Assuming that the utility updates rates relatively fast compared with how customers switch, it is possible to show a result akin to Theorem 5 in [15]. The result provides an intuition about the welfare effects of customers switching, including effects on these customers and on the population as a whole.

Proposition 6. *If immediately after a rate update the first group of customers switching leave h_1 to enroll in h_2 , then (i) the surplus of switchers increases, (ii) the difference between surplus of customers in h_1 and h_2 decreases, (iii) the aggregated welfare increases, and (iv) the marginal benefit of the switch decreases.*

The final proposition shows that a regulator could induce the optimum of the Ramsey problem instructing an incentive rule such as the one we propose. Interestingly, this incentive rule does not require the regulator knowing the optimal demand mix in advance. The current conditions fully determine the incentive.

Proposition 7. *The incentive rule that sets rate levels at $(l(\alpha), \bar{p}(\alpha))$, when α describes the demand mix, induces the Ramsey optimum in the long-run.*

Beyond being compelling from a theoretical perspective, this result enhances the consistency of the method introduced in this paper. When using (4.12)–(4.14) to compare structures, a researcher compares among best cases which could be implemented. The result also improves the value of the Ramsey problem as a benchmark case for planning studies.

4.4 Solution Method

In principle, one could solve (4.12)–(4.14) with a non-convex optimization package. Here we describe an alternative approach that exploits Proposition 4. We propose solving (4.15), thereby solving implicitly the Ramsey problem. Because $g(\alpha)$ does not have an explicit functional form, we use an iterative procedure comprising an inner and outer routines. The inner routine solves P_α , a task that, in general, a convex optimization solver can handle. The outer routine searches an optimum for (4.15). At each iteration it uses information of the optimal solution of P_α in order to compute a search direction and a step size. The outer routine can be implemented with a variety of nonlinear optimization algorithms; in this paper we implement it with a *Barrier* method [18].

The barrier method works by dropping the problem's inequality constraints and augmenting the objective with a *barrier function* $\phi(\cdot)$. The new objective is $z_t(\alpha) := tg(\alpha) + \phi(\alpha)$, where t is a weighting parameter that changes across iterations. We define $\phi(\cdot)$ as $\phi(\alpha) := \sum_{i \in I} \ln(\nu_i - \Gamma_{i\bullet}\alpha) + \sum_{h=1}^n \ln(\alpha_h)$, where $\Gamma_{i\bullet}$ is the i th row of the matrix Γ . The new problem is an unconstrained nonlinear program, thus, it can be handled with a Newton-like method. Algorithm 4 describes the outer routine.

Algorithm 4 Outer routine solved via Barrier method

Initialization: Given a strictly feasible α , $t \leftarrow t^0 > 0$, $\mu > 1$ and $\epsilon > 0$
while $(|I| + n)/t \geq \epsilon$ **do**
 $t \leftarrow t + \mu$
 Find with Newton method $\alpha^* := \arg \max_{\alpha} z_t(\alpha)$
 $\alpha(t) \leftarrow \alpha^*$
end while

In practice, one cannot use a standard Newton method in order to maximize $z_t(\cdot)$. Because this function does not have an analytic formula, there is no analytic expression for its inverse Hessian. A suitable approach is using a quasi-Newton algorithm [94]. These methods use an approximation of the inverse Hessian when computing a Newton step. The more sophisticated versions calculate improved approximations of this matrix using first order information gathered as the optimization procedure progresses. In this paper we use the *Limited Memory Broyden-Fletcher-Goldfarb-Shanno* method (L-BFGS), which guarantees R-linear convergence for uniformly concave problems. The performance of this algorithm improves considerably if $\nabla[z_t(\alpha)] = t\nabla g(\alpha) + \nabla\phi(\alpha)$ is available analytically. While deriving an analytic expression for the second term is straightforward, for the first we use a result from sensitivity analysis for nonlinear programs.¹⁰ Its explicit formula is

$$[\nabla g(\alpha)]_h = E \left[\bar{S}^h(\bar{p}^h(\alpha)) - \bar{\lambda}(\alpha)^\top \bar{D}^h(\bar{p}^h(\alpha)) \right] - r_h. \quad (4.17)$$

Performance To test the performance of our approach, we constructed a simple experiment that is similar to the one described in Section 4.5 with two key differences. First, we used only two tariffs: (i) real-time pricing (RTP), for which the volumetric charge varies freely, and (ii) flat rate, for which the volumetric charge is constant across all time steps and outcomes. Second, in addition to examining a scenario with a sample spaces composed of 365 outcomes (as in Section 4.5), we also examined a smaller case with 50 outcomes. For each scenario, we implemented the preceding algorithm, which we will refer to as *parametric*, as well as a *first order* and *second order* procedures. The latter two methods solve the Ramsey problem directly with a nonlinear solver. While the first order passes to the solver the analytic formula of the gradients, the second order provides the analytic Hessian as well. We tested each algorithm-scenario combination for 20 different parameter choices corresponding to a range of technology costs and demand price-responsiveness parameters. The stopping criterion for all algorithms was the same: the lesser of the time to converge within a 10^{-6} duality gap, or 14,400 seconds.

¹⁰For details see the proof of Proposition 2.

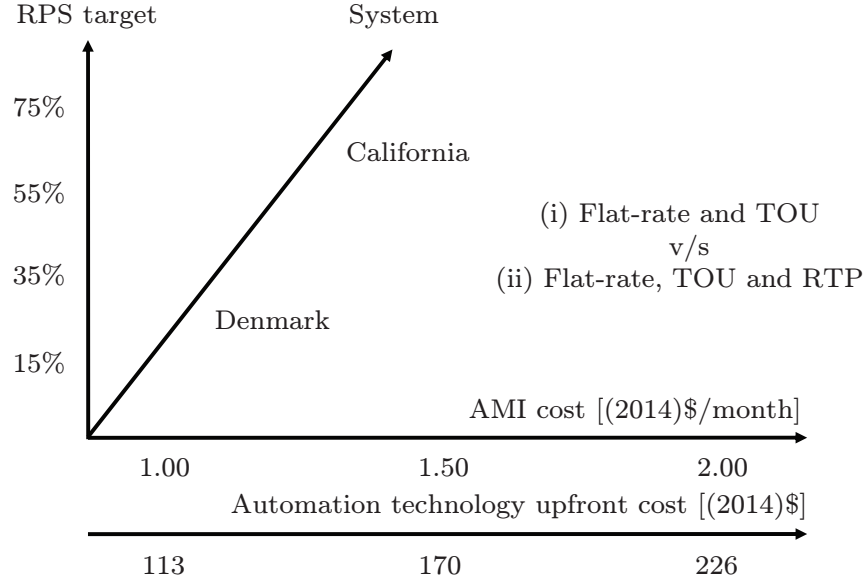


Figure 4.1: Structure of Analysis

Overall, we found that the parametric method converged in 13 ± 5 s (for 50 states of nature, \pm denotes standard deviation) and 226 ± 79 s (365 states of nature). The second order method converged in 49 ± 25 s and $3,132 \pm 1,762$ s (50 and 365 outcomes, respectively) and the first order method converged (or reached the maximum compute time) in $2,865 \pm 2,568$ s and $11,405 \pm 5,050$ s.

4.5 An Application: The Value of Real-Time Pricing

4.5.1 Analysis Design and Data Assumptions

We compared two portfolios of rate structures: (i) a portfolio that includes a flat rate and a time-of-use tariff¹¹ and (ii) a portfolio that adds an RTP tariff to the first portfolio. As Figure 4.1 shows, for each portfolio we considered a range of scenarios for demand-side technology costs, a range of *Renewable Portfolio Standard* (RPS)¹² targets, and load and renewable production data from two systems (Denmark and California; for simplicity we restrict renewable production to wind).

Table 4.1 shows the supply-side technologies we considered and their economic parameters. The only relevant supply-side technical parameter is the availability factor, which we set to 85% for all states of nature for non-wind technologies.¹³ For wind, we use historical system-wide hourly capacity factors for 2014, available from California's Open Access Same-time Information System (OASIS) and the website of the Danish transmission system operator.

On the demand side, we consider one customer type¹⁴ and three arrays of technologies, denoted as Tech 1, 2 or 3: (1) a standard meter, (2) advanced metering infrastructure (AMI) or (3) AMI

¹¹Here we consider a TOU with different volumetric charges for each hour of the day.

¹²An RPS target mandates the utility to produce a fraction of its energy with renewables.

¹³The availability factor from NERC's Generating Availability Data System website.

¹⁴We chose one type due to data availability and to simplify the analysis.

Table 4.1: Economic Parameters of Supply-Side Technologies

		Base-load	Mid-merit	Peak	High-peak	Wind
Capital cost		207	85	27	16	225
Fixed O&M	k\$/MW-yr	69	21	16	11	40
Total fixed		227	106	43	27	265
Fuel		11	27	43	66	0
Variable O&M	\$/MWh	5	11	11	11	0
Total variable		16	38	54	77	0

Notes: Non-wind parameters taken from [40]. Wind costs are the average of those from [47] for California and from [145] for Denmark.

plus automation technology. AMI enables customers to participate in TOU or RTP tariffs, whereas automation technology enables customers to automate the price response of their appliances. We model the latter phenomenon assuming different price elasticities for customers with and without automation. Table 4.2 shows these elasticities whose range is taken from empirical estimates in [55].

Table 4.2: Demand Elasticities

No automation		Automation					
		Low increase		Medium increase		High increase	
own ^a	cross ^b	own	cross	own	cross	own	cross
(0.02)	0.07	(0.04)	0.14	(0.05)	0.20	(0.07)	0.27
(0.04)	0.14	(0.05)	0.20	(0.07)	0.27	(0.08)	0.33
(0.05)	0.20	(0.07)	0.27	(0.08)	0.33	(0.10)	0.40

^a Own-price elasticity.

^b Cross-price elasticity.

Note: Rows provide a range of own- and cross-price elasticities.

Tech 1 costs are normalized to zero; the costs to move to Tech 2 or 3 are incremental. AMI costs are based on U.S. DOE data [42, 44].¹⁵ For automation technology we base costs on currently available advanced programmable controllable thermostats. The highest x-axis values in Fig. 4.1 agree with these incremental cost data; the two additional costs correspond to 25% and 50% reductions.

The baseline levels of consumption correspond to the hourly, system-wide load profile for the year 2014 for both, California and Denmark. While for California this data is publicly available in OASIS, the Danish transmission system operator makes the hourly, system-wide load profile available in its web page.

¹⁵These estimates are for capital and installation costs, net of any benefits to utility operations (e.g. meter reading and back-office staffing). The estimates do not include hypothetical reductions in energy or capacity expansion costs.

4.5.2 Results

As Figure 4.2 shows, welfare in Portfolio (ii) exceeds Portfolio (i) across the range of factors we explored. Additionally, higher elasticity levels increase the positive impact of adding RTP to (i). These findings are not surprising in light of previous work (e.g. [15], [40], [38] and [73]). However, our framework allows a more comprehensive analysis.

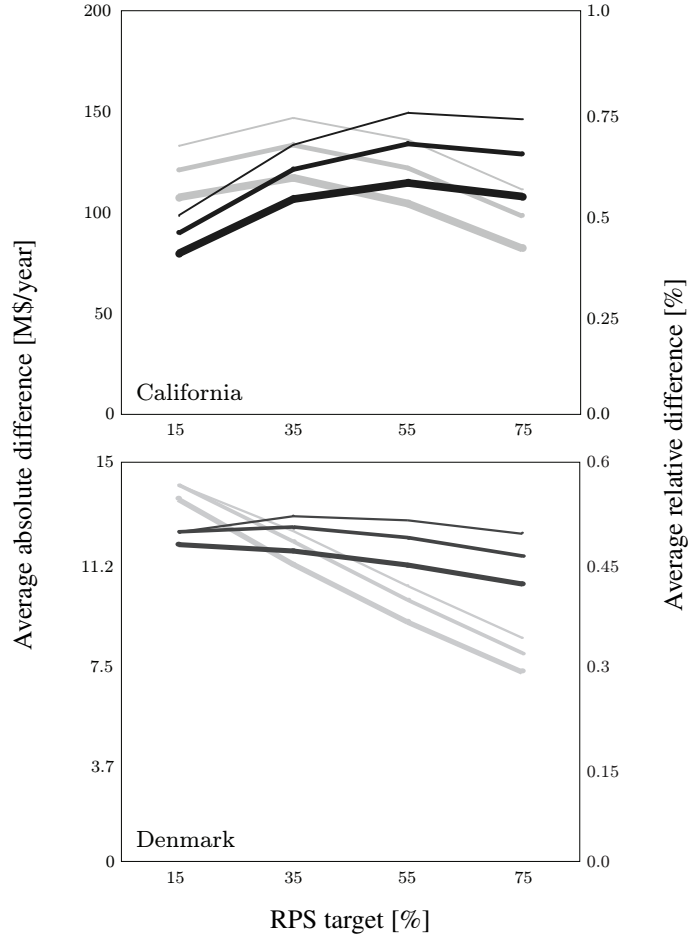


Figure 4.2: Average Welfare Differences Between Portfolios ((ii) - (i)) by RPS Target. Absolute Differences in Black. Relative Differences in Gray. Line Thickness Represents Different Elasticity Levels. Thicker Lines Correspond to Lower Own- and Cross-Price Elasticities.

In terms of relative welfare differences, Portfolio (ii) does not seem particularly attractive. Considering additional rate design factors such as simplicity or public acceptance, a regulator could conclude that neither in California nor in Denmark real-time pricing is valuable enough to justify its deployment. However, a significantly different picture emerges when one observes the results in Table 4.3.

The demand mix provides a complementary perspective on the relevance of demand alternatives for a system. While in California between 13 to 23 percent of the population would enroll in real-time pricing were it present, only a 4 to 5 percent would in Denmark. In equilibrium these

Table 4.3: Demand Mix by RPS Target

			California				Denmark			
			RPS [%]							
Portfolio	Tech	Tariff	15	35	55	75	15	35	55	75
(i)	1	FR	97	97	96	97	96	97	99	99
	2	TOU	2	2	3	2	3	2	1	1
	3	TOU	1	1	1	1	1	1	0	0
(ii)	1	FR	87	82	79	77	96	95	95	95
	2	RTP	9	11	13	14	2	3	3	3
	3	RTP	4	7	8	9	2	2	2	2

Note: Values are percentages of population under each category.

differences appear not relevant because welfare results are similar, in relative terms. But, in view of Proposition 6, the roll-out of RTP will certainly accrue more benefits across time for customers in systems like California's than for those in systems like Denmark's. This policy-relevant perspective cannot be achieved without a modeling framework such as that developed in this paper.

An alternative way of inferring differences in the demand mix would be analyzing net-load duration curves.¹⁶ Since demand responsiveness competes with peaking plants [15, 40], differences during peak-hours should translate into differences in the demand mix. Figure 4.3 shows, however, that small variations in the shape of the curves can imply significantly different demand mixes. Thus, using net-load duration curves in order to anticipate possible differences does not seem a suitable approach.

Finally, we point out two additional conclusions an analyst can derive from the demand mix. First, it allows simplifying portfolios of rate structures. For instance, Table 4.3 shows that when faced with Portfolio (ii) no customer enrolls in the time-of-use program, which indicates that a simpler portfolio achieves the same benefits. Second, the demand mix establishes targets for rolling-out demand technologies and rate structures. These targets are not trivial as they correspond to fractions of the population and are contingent on the many factors we explored. In particular, considering that the AMI costs we used are net of any benefits to utility operations (e.g. meter reading and back-office staffing), our results suggest that 100% smart meter rollouts are not cost-effective in the regions we investigated.

4.6 Conclusions

This work introduces an analytic method for helping planners and regulators in the design of rates in the electricity sector. It develops a nonlinear program that serves as tool to compare portfolios of rate structures, and proposes a suitable approach to find an optimal solution of this model.

¹⁶Net load-duration curves are analysis tools used in the electricity sector to estimate the long term value of different production technologies. They plot hourly net loads (system demand minus the renewable production) for each hour of a period, say a year. Hours are sorted from left to right with the highest net load hour to the extreme left and the lowest to the extreme right. A point in this curve indicates the fraction of the time (x -coordinate) the net load is greater or equal to the net load in the curve (the ordinate). For more details, we refer the reader to [136].

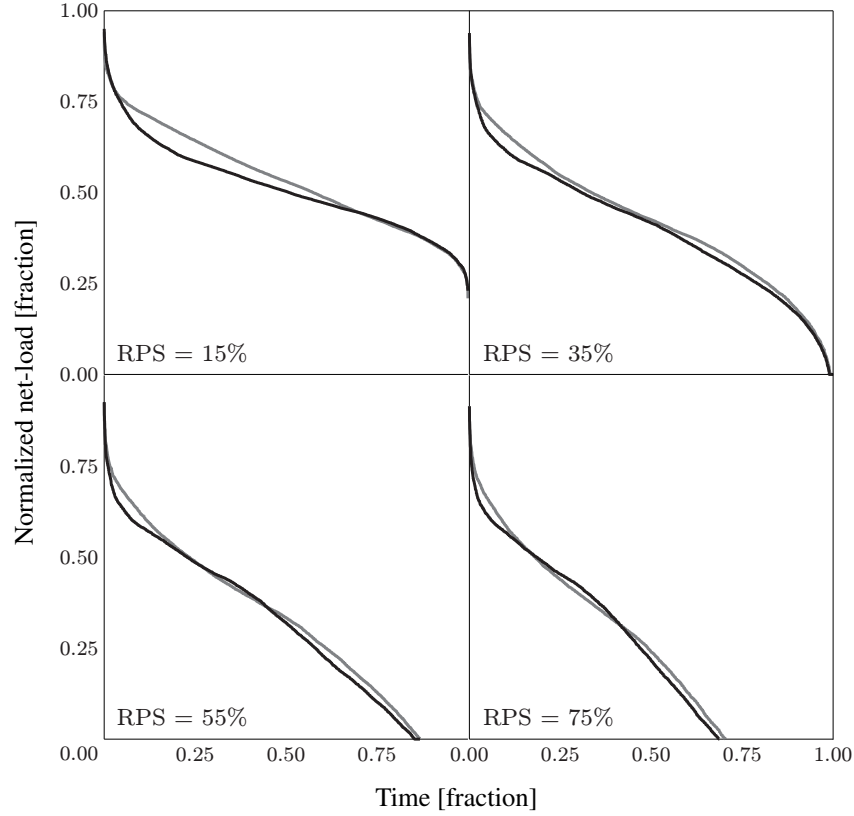


Figure 4.3: Net-Load Duration Curves by RPS Target. California Net-Load in Black. Denmark Net-Load in Gray. All Curves Reach a Maximum of 100%.

The flexibility of our method allows consistently comparing tariff structures, and it enables researchers to model a richer set of trade-offs influencing the production costs in the sector. It also allows the comparison of portfolios of tariffs under optimal demand mixes. A theoretical exploration of the properties of the nonlinear program suggests that our method compares rates which are not only socially optimal but could also be implemented. Besides, the demand mix that the model computes offers a valuable perspective on the potential of competing demand alternatives. It helps planners and regulators to prioritize demand technologies and rates, and to establish appropriate levels of deployment for each demand option.

The application of the framework to the comparison of portfolios of rate structures in California and Denmark shows its practical value. As the theory predicts, real-time pricing increases welfare in both systems. But these benefits may not be enough to deploy it in either. The systems have, however, different demand mixes which indicate different policy prescriptions. While in Denmark RTP appears unattractive, it at least deserves further revision in California.

5. A Mathematical Programming Approach to Utility Pricing

5.1 Introduction

The distribution segment of the supply chain of commodities such as gas or electricity has characteristics of natural monopoly. In these instances one firm, a distribution utility, serves all customers and maintains the entire distribution infrastructure. Absent competition, a regulator supervises the pricing, or rate design, of the utility's services. The resulting regulated prices serve two main purposes. They guarantee that the firm recovers its costs, and thereby sustains its operations, and sends proper economic signals to retail customers [78].

Changes in the landscape of the utility industry are challenging prevailing rate designs. In some sectors, such as in the electricity industry, innovations in information and automation technologies enable utilities to implement more complex tariffs. For instance, advanced metering infrastructure (AMI) allows measuring and recording electricity consumption at the hourly timescale. With this technology utilities can implement time-varying rate structures, designs in which volumetric charges can change across time [77].

The increasing penetration of distributed energy resources (DERs) and intermittent generation also incentivize innovation in rate design. The massive adoption of DERs, such as rooftop solar photovoltaic (PV) panels, home energy management systems (HEMS) or electric vehicles (EVs), has pushed regulators to rethink the way in which utilities should collect their revenue [105]. In a world of pervasive DERs, innovative tariffs design are expected to improve the efficiency of the sector, decreasing short- and long-run systems costs, and tackling distortions such as the cross-subsidization between customers with and without DERs [105].

On the other hand, high-voltage grid-connected intermittent generation technologies, such as wind-mills or solar photovoltaic systems, bring new challenges to the grid operations. The output of these technologies is driven mainly by weather conditions, such as wind speed, wind direction, cloud cover or haze, which change considerably across time. As a result, systems operators cannot dispatch these resources at will. In absence of economically feasible storage, the intermittency of wind or solar requires additional reserves to ensure reliability and more capacity to meet demand. These requisites could translate into higher production costs and even undermine carbon emission reductions [59]. Rate structures that reflect closely system conditions could work in sync with renewable resources, incentivizing consumption when these resources are available, and bringing demand down when they are scarce.

In theory, there is one type of rate that could materialize all these benefits: a two-part real-time pricing (RTP). This variant of a time-varying rate, in which volumetric charges can differ from hour to hour, outperforms all other structures from an efficiency perspective [73]. Despite of this fact, at the time of this writing, only a small number of jurisdictions have implemented this structure [131]. Part of the reason is that designing rates not only involves economic considerations. Regulators must balance other objectives such as the simplicity, distributional impacts and stability of competing tariff designs [10]. Having techniques that allow consistently quantifying the economic differences of various rates brings more clarity to the overall analysis. Regulators can balance economic gains of more efficient rates, such as a real-time pricing, versus other objectives. Even though RTP could maximize efficiency, the economic gains with respect to a more simple or

stable rate could be small enough to advise against its implementation.

This paper contributes with a quantitative technique to evaluate rate structures. In contrast to previous approaches, our method allows comparing within a unified framework a large class of tariff designs. This includes time-varying structures, rates with demand charges, or charges for peak-consumption, and block rates—tariff with volumetric charges contingent on total consumption. A second distinctive characteristic of the present technique is that it enables modeling in a transparent manner important aspects emerging in the utility industry. Our framework allows representing in detail distributed energy resources, such as PV panels or battery storage systems, and flexible appliances, such as heating and cooling systems. This realistic representation of the demand side is embedded within a traditional capacity expansion setting (e.g., [104]). This allows researchers to explore the long-run implications of different rate structures, taking into account their interaction with the full supply chain. More importantly, it enables to model the time-variant and stochastic nature of intermittent technologies, resources which are increasingly important in modern utility sectors.

Our model builds upon the theory of Peak-Load Pricing, a framework that captures the interaction between rate structures and investment decisions in the utility sector, and can be used to compare rate designs. In this setting, a regulator chooses the pricing of a monopolistic utility in order to maximize societal welfare, solving what is called the Ramsey-Boiteux problem [76]. In doing so, the regulator, or Ramsey planner, internalizes the consumer responses to different prices and the cost function of the monopolist, which together determine the optimal allocation [38]. One can use this framework to compare rate structures. By solving the model with different constraints sets for the prices, researchers can compare the characteristics of the resulting equilibria.

The Ramsey-Boiteux problem is an instance of a Bilevel Model, a type of mathematical programming problem in which group of variables is constrained to be in the solution sets of subordinate, or lower level, problems [41]. One way of approaching a Bilevel Problem is replacing these solution sets by the first order necessary conditions of the lower level problems. The resulting model is a Mathematical Program with Equilibrium Constraints (MPEC), which can be handled with specialized nonlinear algorithms.

Because of dimension of the type of problems we attempt to tackle with the present technique, we develop a decomposition approach for the MPEC version of the Ramsey-Boiteux problem (RMPEC). The algorithm is a nonlinear variation of the Alternating Directions Methods of Multipliers (ADMM), a distributed computation technique which blends the decomposability of dual subgradient methods and the convergence properties of the method of multipliers. ADMM algorithms solve optimization problems via successive iterations, each of which optimizes an augmented problem along two blocks of variables. Typically, the resolution of one block can be distributed, and the other uses as input parameters the distributed solutions. In our implementation, the distributed step involves solving low-dimensional MPECs, that we tackle using the mixed integer programming representation of [58]. To handle the problem of the other block of variables we use conic programming. While inexact, our variant of ADMM allows tackling large-scale instances of the Ramsey-Boiteux problem, greatly enhancing the applicability of our approach.

With a computational experiment, we explore the performance of the present technique. We test the algorithm on 200 hundred instances, which we construct varying the number of variables and

parameters of the model. We find that the algorithm has desirable properties for practical applications. It vastly outperforms a popular commercial solver for Mathematical Programs with Equilibrium Constraints: Knitro. While this package is not able to find a solution within 24 hours even for instances of small size, our algorithm converges within 2 hours in 94% of the instances tested. Furthermore, the results suggest that the algorithm is suitable for distributed computation. The distributed step of our variant of ADMM increases close to linear with the size of the problem while the centralized step grows at a rate lower than linear.

In order to illustrate the value of the technique as a tool to compare rates in the utility sector, we conduct an analysis of tariffs structures in a simplified setting. The analysis highlights the value of the modeling flexibility that the present technique provides. It shows how abstracting from network constraints can introduce significant distortions in the analysis of rates. In our exercise, omitting the existence of a network leads to underestimating the benefits of time-varying rate structures by close to 31 times. In addition, the analysis shows that DERs could potentially complement time-varying rates, increasing the value of this structures while at the same time these type of rates could incentivize greater adoption of distributed resources. Finally, the exercise highlights the value of being able to explore impacts of rate design on a population of heterogeneous customers. While in all cases switching from a flat rate to a more sophisticated rate structure improves the welfare of the aggregate of households, some rates benefit wealthier customers more. Moreover, there are time-varying structures that can even harm customers with low levels of wealth, being these worse off after the switch.

5.2 Pricing Utility Services

The pricing of utility services involves the determination of a *revenue requirement* and *rate structures*. While the former corresponds to the total compensation utilities receive from its customers, the latter are the instruments they use to collect such compensation.¹ To the extent that rate structures have an impact on consumption behavior which affects operational and capital expenses, determining the revenue requirement and the structure of the rates are not independent efforts. However, it has been the industry and academic practice to consider these steps independently, understanding the complexities involved in one task while simplifying the other (see, e.g., [76], [120]). In this work we adhere to this strategy, focusing on the determination of rate structures.

One can further divide the definition of a rate structure into selecting the set of charges that the tariff will include—what we call its design—and setting the actual values of these charges. The technique we contribute with allows regulators and policy makers to discern among competing rate designs. For simplicity and because it is the focus of this paper, we refer to the design of a rate as rate structure or tariff structure as well.

5.2.1 Quantitative Methods for Evaluating Rate Structures

Applied analyses of rate structures in the utility sector have sought measuring welfare changes resulting from modifications in the design of the prevailing rates. The techniques developed in

¹For a more complete discussion on the subject of rate design, see [76]

these studies differ in how they quantify a change on welfare, ΔW . All make use of the following identity,

$$\Delta W = \Delta Y + \Delta \Pi - \Delta r, \quad (5.1)$$

where Δr correspond to a change in customer related costs (e.g., capital cost of AMI), $\Delta \Pi$ change in production costs and ΔY is the compensating variation—the money that when taken away from individuals leaves them with the same level of welfare they had before the price change [98]. Based on how the studies treat ΔY , we distinguish two groups. [30], [60], [69], [91] and [111] start assuming a functional form for the indirect utility function—the optimal value of the consumer utility maximization problem, estimate the parameters of a theoretically consistent demand model, and use these estimates plus the relationship between the indirect utility and compensating variation to derive the latter.² A second group uses the consumer surplus to approximate ΔY . [2], [12], [15], [139] and [4] start by assuming a functional form for a system of demand equations, estimate or calibrate its parameters, and integrate the system in order to compute the consumer surplus. [153] shows that when focusing on goods with an associated expenditure relatively small with respect to the customer’s budget—such as the consumption of utility services, consumer surplus and compensating variation are equivalent.

The method that we propose in this paper falls within the second strand of approaches and extends previous techniques in significant ways. It allows the consistent comparison of a wide range of rate designs, including time-varying rate structures, rates with demand charges and increasing or decreasing block rates. With the exception of [60], previous studies have developed techniques to measure welfare differences between variations of time-varying rate designs, such as a flat rate (FR), a time-of-use rate (TOU) or a real-time pricing structure (RTP). These approaches have either compared tariff structures specifying ex-ante the value of the volumetric charges, or imposing unnecessary constraints on them (see, e.g., [2] or [12]). [60], on the other hand, develops a technique that permits researchers to compare time-varying rate structures with designs of the same kind supplemented with demand charges. In their setting volumetric charges are endogenous. This approach, however, does not permit to model other important groups of rate design such as block rate structures.

A second limitation of [60], which is shared by all other techniques with the exception of [30], is a simplified representation of the supply side. Previous work has either ignored the supply side (e.g. [4], [69] and [139]), or has simplified its representation with a cost function (e.g. [15] and [60]). Even though, [30] take a more comprehensive approach, using a detailed economic dispatch model, they use this model only to estimate a marginal production cost function. Our technique, on the other hand, finds rates while simultaneously optimizing short- and long-run production decisions for a wide variety of production technologies and in the presence of key infrastructure such as a transmission network.

While we believe these contributions are valuable, the most salient aspect of the present technique is that it allows researchers to construct a bottom up model of customer behavior. In our setting, the demand of a customer is the solution of a utility maximization problem subject to a set of constraints. This constraints can be defined by the researcher providing the flexibility to model a

²For an individual with an income I , the relationship between the compensating variation and the indirect utility function when prices change from p^0 to p^1 is $v(p^0, I) = v(p^1, I - \Delta Y)$, where $v(\cdot, \cdot)$ is the indirect utility function [98].

wide variety of devices, such as HVACs systems or refrigerators, or distributed energy resources, such as rooftop solar PV or battery storage systems. This modeling flexibility is fully unique to our method; all previous techniques use a top-down, aggregated representation of the demand. A bottom up model overcomes three important limitations of past approaches. One is that these implicitly consider a population of homogeneous customers. While heterogeneity may not be crucial in determining the aggregated benefits of a change in the design of rates, it will certainly help to understand the implications for different types of customers. One cannot use a model with a representative customer to study distributional impacts. A second drawback is that a top down approach precludes researchers from including in a transparent manner important determinants of consumption behavior, such as weather patterns. The third limitation, which is specially relevant today, is the impossibility of using top down approaches to study distributed energy resources. By their very nature, these technologies are sources of heterogeneity in the population, geographically, and in terms of consumption patterns. Given that rate design can not only influence how households use these resources but also how they adopt them, it is important to have tools that permit anticipate plausible outcomes.

We build our model considering as starting point the model we developed in the previous chapter, which is based on the theory of Peak-Load Pricing. For completeness and to introduce the notation that we will use throughout the chapter, we now succinctly review this theory.

5.2.2 Peak-Load Pricing: A Theoretical Framework to Compare Rate Structures

Let Ω be a discrete sample space, q_ω the probability that $\omega \in \Omega$ occurs, and $E[\cdot]$ the associated expectation operator. We refer to an element in Ω as outcome or state of nature, and distinguish a random from a deterministic variable writing the former in boldface. Given a random variable \mathbf{y} , we denote y_ω the realization of this variable when ω occurs. The symbol \top indicates the transpose of a vector.

The theory of Peak-Load Pricing has as objective to provide guidelines for the pricing of public utility services [38]. Its starting point is the problem faced by a regulator that set prices with the aim of maximizing societal welfare. At the same time, the regulatory body must guarantee that the regulated monopolist is able to cover its costs.

The monopolist serves a population of customers with different types $i \in I$. These distribute in the population according to the frequency function $\delta(\cdot)$, such that $\delta(i)$ is the number of customers with type i . The monopolist offers a set $\{1, \dots, T\}$ of commodities and the customers decide among consumption bundles $d \in \mathbb{R}_+^T$. A quasi-linear utility $U(d; \theta_\omega^i) + m_\omega$ characterizes the preferences of types i over these bundles, with $U : \mathbb{R}_+^T \rightarrow \mathbb{R}$ and θ_ω^i a set of exogenous parameters. A type i has limited budget M_i . The customer's demand for each commodity results from her choosing optimally among bundles, i.e.,

$$D(p_\omega; \theta_\omega^i) := \arg \max_{d \geq 0} \{U(d; \theta_\omega^i) + M_i - p_\omega^\top d\}. \quad (5.2)$$

The theory assumes that U is strictly concave so that $D(p_\omega; \theta_\omega^i)$ is a singleton; and the gross surplus of this customer is $S(p_\omega; \theta_\omega^i) := U(D(p_\omega; \theta_\omega^i))$. In addition for $I' \subseteq I$, we denote $\mathbf{D}^{I'}(\mathbf{p}) :=$

$\int_{I'} D(\mathbf{p}; \boldsymbol{\theta}^i) \delta(i) di$ the aggregated demand of the types in I' , define $\mathbf{S}^{I'}(\mathbf{p})$ in a similar fashion, and call $\nu_{I'} := \int_{I'} \delta(i) di$ the total number of customers with types in I' .

The monopolist collects its revenue with a two-part rate structure, a contract (l, \mathbf{p}) where l is a fixed (or customer) charge and \mathbf{p} a vector of charges per unit of consumption. The corresponding consumer surplus is

$$CS(l, \mathbf{p}) = E [\mathbf{S}^I(\mathbf{p}) - \mathbf{p}^\top \mathbf{D}^I(\mathbf{p})] - l \cdot \nu_I. \quad (5.3)$$

There is a set of production technologies that we index with the letter $k \in K$. Each technology differs from others on its variable costs per unit of production, $c_{\omega k} \in \mathbb{R}_+^T$, its fixed costs \hat{r}_k , and its availability factor, $\rho_{\omega k} \in \mathbb{R}_+^T$. The latter captures the variability in the technology's availability due to, for instance, the intermittent output of some renewables or the occurrence of outages. The installed capacity of technology k is x_k and $y_{\omega k} \in \mathbb{R}_+^T$ is its production vector. With this definitions, the production cost for a bundle \mathbf{d} is

$$C(\mathbf{d}) = \min_{(x, y)} \sum_{k \in K} E [\mathbf{y}_k^\top \mathbf{c}_k + x_k \hat{r}_k] \quad (5.4)$$

subject to

$$\mathbf{d} \leq \sum_{k \in K} \mathbf{y}_k, \quad (5.5)$$

$$0 \leq \mathbf{y}_k \leq x_k \boldsymbol{\rho}_k, \quad k \in K \quad (5.6)$$

The profit of the monopolist is

$$\Pi(l, \mathbf{p}) = E [\mathbf{p}^\top \mathbf{D}^I(\mathbf{p})] + l \cdot \nu_I - C(\mathbf{D}^I(\mathbf{p})) - \Pi_0, \quad (5.7)$$

where Π_0 captures transmission and distribution costs, overhead expenses and the opportunity cost of the monopolist. We note that this profit function can also represent the aggregated profits of a sector in which there is perfect competition at the wholesale level, and there is a regulated utility at the retail level (see [73], [32]).

The welfare maximization problem or simply the Ramsey problem is

$$\max_{(l, \mathbf{p})} \{CS(l, \mathbf{p}) : \Pi(l, \mathbf{p}) \geq 0, (l, \mathbf{p}) \in \mathcal{L} \times \mathcal{P}\}. \quad (5.8)$$

Henceforth we refer to $\mathcal{L} \times \mathcal{P}$ as *rate structure*, and to an element of this set as *rate level*.

5.3 An Alternative Quantitative Technique

The model we developed in the previous chapter extends the basic setting Peak-Load Pricing. We now briefly review this model and discuss its limitations.

Let $h \in \{1, \dots, H\}$ index the rates in the portfolio, and redefine $\mathcal{P} := \mathcal{P}_1 \times \dots \times \mathcal{P}_H$, and $\mathcal{L} := \mathcal{L}_1 \times \dots \times \mathcal{L}_H$ such that \mathcal{P}_h is the feasible region for the volumetric charges of rate h and \mathcal{L}_h constraints the fixed charges of the same rate. In this setting h also indicates the DER a customer adopts. The letter α_h denotes the number of customers enrolled in h , and the vector $\alpha \in \mathbb{R}_+^H$ represents the distribution of the population across rates, which we call demand mix. In

this context I is a discrete set indexed by i , with ν_i being the number of customers with type i . Defining the matrix Γ such that $[\Gamma]_{ih} = 1$ if a type i can enroll in h and 0 otherwise, the feasible region for α is $\mathcal{A} := \{\alpha \in \mathbf{R}_+^n : \Gamma\alpha \leq \nu\}$. As in the basic Peak-Load Pricing framework, the functions U, D, S correspond to the direct utility, demand and gross surplus, respectively; and a set of endogenous parameters θ^h determines them for each h .

The consumer surplus function in this model is

$$CS(l, \mathbf{p}) = \sum_{h=1}^n \alpha_h E [S(\mathbf{p}^h; \theta^h) - D(\mathbf{p}^h; \theta^h)^\top \mathbf{p}^h] - l^\top \alpha, \quad (5.9)$$

and the surplus of the regulated utility

$$\Pi(l, \mathbf{p}) = \sum_{h=1}^n \alpha_h E [D(\mathbf{p}^h; \theta^h)^\top \mathbf{p}^h] + (l + r)^\top \alpha - C \left(\sum_{h=1}^n \alpha_h D(\mathbf{p}^h; \theta^h) \right). \quad (5.10)$$

Because $\Pi(l, \mathbf{p}) = 0$ in the optimum, we can write the Ramsey problem as follows

$$\max_{(\alpha, \mathbf{p})} \sum_{h=1}^n \alpha_h E [S(\mathbf{p}^h; \theta^h) - r_h] - C \left(\sum_{h=1}^n \alpha_h D(\mathbf{p}^h; \theta^h) \right) \quad (5.11)$$

subject to

$$\mathbf{p} \in \mathcal{P}, \quad (5.12)$$

$$\alpha \in \mathcal{A}. \quad (5.13)$$

5.3.1 Limitations of the Model

The Ramsey problem has some limitations as a model to compare rate structures. The main two relate to the realism of the demand representation and the type of rates that could be modeled with this framework. The demand representation in (5.11)–(5.13) is similar to that of the basic Peak-Load Pricing model in that it is the solution of a utility maximization problem akin to (5.2). The theory makes the assumptions needed so the solution set of this problem is a singleton. While this simplification makes the Ramsey problem amenable to mathematical analysis, it comes along with two mayor drawbacks. First, in many cases of interest, the solution set of the utility maximization problem may not be a singleton. Example 5.3.1 shows a case of great relevance in modern electricity systems.

Example 5.3.1. Consider a household with an electric vehicle (EV) enrolled in a flat rate with volumetric charge $p \in \mathbb{R}_+$. For simplicity, we consider that the round-trip efficiency of the vehicle's battery is 1. While implausible, the reader can verify that this assumption does not alter the point we make with this example. The aggregated demand of the household is $d + s$, where d is the electricity consumption of the household and s that of the EV. The battery of the electrical vehicle has a maximum and minimum charge and discharge rates $R^+ > R^-$, and a maximum and minimum

state of charge $E^+ > E^-$. Consistently, the feasible region for s is

$$\mathcal{S} = \left\{ s \in \mathbb{R}^T : s_0 + \sum_{\tau=1}^t s_\tau \in [E^-, E^+] \wedge s_t \in [R^-, R^+] \quad \forall t \right\}, \quad (5.14)$$

with s_0 the initial state of charge. The utility maximization problem is

$$\max_{(d,s)} \left\{ U(d; \theta) + M - p \sum_{t=1}^T (d_t + s_t) : d \geq 0, s \in \mathcal{S} \right\} \quad (5.15)$$

Let s^* be optimal for (5.15), suppose that there is a pair of consecutive periods where $s_t^* < s_{t+1}^*$, and define $\Delta := (s_{t+1}^* - s_t^*) \cdot \psi$. For any $\psi \in (0, 1)$, we have that $s^{**} = (s_1^*, \dots, s_t^* + \Delta, s_{t+1}^* - \Delta, \dots, s_T^*)$ is also optimal for (5.15). \square

A second problem with the demand representation of the Ramsey problems relates to the calibration of this function. Researchers have estimated price elasticities for electricity demand. One can classify the approaches in two groups. One group focuses on estimating own-price elasticities and peak-to-off-peak elasticities (e.g., [30], [69] and [91]). While these have been useful for analysis of time-off-use rate structures, they impose limits in terms of the range of rates that a researcher can analyze. To evaluate rate structures such as a real-time pricing, the approaches used in these papers fall short, as peak-to-off peak substitution is not a relevant concept in the case of RTP. To overcome these limitations other researchers have used techniques that permit to compute a full matrix of elasticities (e.g., [60] and [139]). While this approach brings more flexibility to the analysis of rates, it has three limitations. One is that the demand system must be a linear function of the price, which limits the range of utility maximization problems one can consider in an analysis. A second drawback is the lack of transparency when introducing heterogeneity. In principle, one can capture heterogeneity estimating various elasticity matrices. Having these inputs, however, makes it hard to understand the role of fundamentals, such as weather patterns or new technological conditions, on the results of an analysis. A third limitation is range of rates that researchers can model. This framework permit analyzing any variations of a time-varying rate structures. However, other tariffs such as increasing block rates or rate structures supplemented with demand charges cannot be modeled.

5.3.2 A Realistic Demand Model

We believe that a bottom up approach can tackle these limitations. It can simplify and make more transparent the calibration of the demand model. Further, a bottom up approach can enable the analysis of more sophisticated rate structures. We slightly modify the framework of the previous chapter to allow the bottom up modeling of the demand. Instead of considering the demand function as a primitive, we consider as fundamental inputs the elements defining the consumer maximization problem. In the present extension, the demand is simply some optimizer of this problem. The new version of the Ramsey problem follows

$$\max_{(\alpha, \mathbf{d}, \mathbf{p})} \sum_{h=1}^n \alpha_h E[U(\mathbf{d}^h; \boldsymbol{\theta}^h)] - C \left(\sum_{h=1}^n \alpha_h \Psi_h \mathbf{d}^h \right) \quad (5.16)$$

subject to

$$\mathbf{d}^h \in \arg \max_{\mathbf{d}} \{E[U(\mathbf{d}; \boldsymbol{\theta}^h) + M_h - \mathbf{d}^\top \Lambda^h \mathbf{p}^h] : b^h - A^h \mathbf{d} \geq 0\}, \forall h \quad (5.17)$$

$$\mathbf{p} \in \mathcal{P}, \quad (5.18)$$

$$\alpha \in \mathcal{A}, \quad (5.19)$$

where for simplicity we have added the parameter r_h to the set of parameters $\boldsymbol{\theta}^h$. In this formulation, Λ^h, b^h, A^h permit modeling customers with a wide array of DERs as well as more complex rate structures. The parameter Ψ_h is a demand aggregation matrix. For concreteness, we now provide illustrative examples.

5.3.3 Examples

We start showing how to set the parameters Λ^h, b^h, A^h and Ψ_h to model a customer having DERs and enrolled in a simple flat rate structure. Next, we show how the parameters change to model the same customer under more complex rates. In both examples, we drop the subindex h since we focus on one customer.

Example 5.3.2. Consider a household with a photovoltaic solar panel (PV), a battery storage systems (BS), and a thermostatically controlled load (TCL)—the latter regulates the temperature inside the customer premises (e.g., an AC system). Define $J = \{\text{TCL}, \text{PV}, \text{BS}, \text{OD}\}$ as the of devices the household owns, with OD representing all other devices of this customer. The demand vector of the household is $\mathbf{d} = (d_{TCL}, d_{PV}, d_{BS}, d_{OD})$, and the customer's consumption choices are consistent with an additively separable utility function $U(\mathbf{d}) = \sum_{j \in J} U_j(d_j)$. Neither the electricity consumption of the PV nor that of the BS produce any benefit for the household so $U_{PV}(d_{PV}) = U_{BS}(d_{BS}) = 0$. We leave the utility function associated to other devices (OD) as generic and concentrate in specifying the one corresponding to the TCL. The inside temperature of the household w can be modeled as a linear function of the power consumption of the TCL

$$w(d_{TCL}; \xi, \hat{w}) = W_1(\xi) d_{TCL} + w_2(\xi, \hat{w}), \quad (5.20)$$

where ξ is the set of thermal characteristics of the dwelling and \hat{w} is the outdoor temperature; W_1 and w_2 are a matrix and a vector depending on these parameters. We specify the utility function associated to the TCL as the negative of the disutility that deviations from a desired indoor temperature, w_{target} , produce on the customer. That is,

$$U_{TCL}(d_{TCL}) = -\beta \|w(d_{TCL}; \xi, \hat{w}) - w_{target}\|^2, \quad (5.21)$$

where β is a parameter characterizing how the household trades comfort for savings.

Let \mathcal{S}_j be the constraint associated to the end use j . We have already defined \mathcal{S}_{BS} in (5.14).

Since this set is a polyhedron, we can write $\mathcal{S}_{BS} = \{d \in \mathbb{R}^T : b_{BS} - A_{BS}d \geq 0\}$. The constraint set for the photovoltaic system is simply $\mathcal{S}_{PV} = \{d \in \mathbb{R}^T : d_t \in [0, x\rho_t] \forall t\}$, with x and ρ the nameplate capacity and availability factor of the PV, respectively. Again, \mathcal{S}_{PV} is a polyhedron so it can be written as the intersection of halfspaces $b_{PV} - A_{PV}d \geq 0$. The final group of technological constraints corresponds to those imposing power limits one the TCL. Let $\mathcal{S}_{TCL} = \{d \in \mathbb{R}^T : d_t \in [0, d_{max}] \forall t\}$; the corresponding inequality is $b_{TCL} - A_{TCL}d_{TCL} \geq 0$. The full set of constraints for the household demand follows

$$\underbrace{\begin{bmatrix} b_{TCL} \\ b_{PV} \\ b_{BS} \\ z \end{bmatrix}}_b - \underbrace{\begin{bmatrix} A_{TCL} & Z & Z & Z \\ Z & A_{PV} & Z & Z \\ Z & Z & A_{BS} & Z \\ Z & Z & Z & -I_d \end{bmatrix}}_A \begin{bmatrix} d_{TCL} \\ d_{PV} \\ d_{BS} \\ d_{OD} \end{bmatrix} \geq 0, \quad (5.22)$$

where Z and z are, respectively, a matrix and a vector of zeros of the proper dimensions and I_d is an identity matrix. The corresponding demand aggregation matrix is $\Psi = [I_d \ -I_d \ I_d \ I_d]$.

Given that the household is under a flat rate, $\mathcal{P} = \{p \in \mathbb{R}^T : p_{\omega t} = p_{\omega' t'} \forall (\omega', t')\}$ and $\Lambda = \Psi^\top$. \square

In the example that follows, we consider the same household enrolled in a flat rate structure supplemented with a demand charge—a charge per unit of peak consumption.

Example 5.3.3. Redefine the households demand as follows $d \leftarrow (d, d_{DC})$, where $d_{DC} \in \mathbb{R}$ is the consumption on peak. We also update the parameters of the constraint (5.22) assigning

$$b \leftarrow \begin{bmatrix} b \\ z \end{bmatrix} \quad \text{and} \quad A \leftarrow \begin{bmatrix} A & z \\ \Psi & -e \end{bmatrix}, \quad (5.23)$$

with e a vector of ones of dimension T . The price-demand multiplication matrix $\Lambda \leftarrow [\Lambda \ z; \ z^\top \ 1]$ and the new demand aggregation matrix is now $\Psi \leftarrow [\Psi \ z]$. Finally, the constraint for the prices updates as follows $\mathcal{P} \leftarrow \mathcal{P} \times \mathbb{R}_+$. \square

In our last example we also consider as starting point the definition of the parameters and variables in Example 5.3.2. We show how to update such parameters and variables in order to model a customer enrolled in a increasing block (IB) structure.

Example 5.3.4. In the case of an IB design a customer pays a volumetric charge which differs depending on the level of her total consumption over a certain horizon. Here we consider that the relevant horizon is $\{1, \dots, T\}$. There are N consumption blocks with upper bounds $\{q_n\}_{n=1}^N$, corresponding to the components of the vector $q \in \mathbb{R}_+^N$. In an increasing block structure if the total consumption is within block n , i.e., if it falls in the interval $[q_{n-1}, q_n]$, then the per unit charge is at least as high as that of the block n' , for any $n' < n$.

We start redefining the household demand $d \leftarrow (d, d_{IB})$, where $d_{IB} \in \mathbb{R}^N$ has in its n -th component the total consumption if it falls in the n -th block. The new parameters of the constraints

follow

$$b \leftarrow \begin{bmatrix} b \\ 0 \\ q \\ z_N \end{bmatrix} \quad \text{and} \quad A \leftarrow \begin{bmatrix} A & Z \\ e_T^\top \Psi & -e_N^\top \\ Z' & I_N \\ Z' & -I_N \end{bmatrix}, \quad (5.24)$$

with Z, Z' matrices of zeros of the proper dimensions, z_N an N -dimensional zeros vector, e_T and e_N vectors of ones with T and N components, respectively, and I_N the identity of dimension N . The price-demand multiplication matrix changes to $\Lambda \leftarrow [Z' \ I_N]^\top$, and the demand aggregation matrix becomes $\Psi \leftarrow [\Psi \ I_N \cdot 0]$. Finally, we redefine the constraint set for the price vector as

$$\mathcal{P} \leftarrow \{p \in \mathbb{R}_+^N : p_n \geq p_{n-1} \ \forall n \in \{2, \dots, N\}\}. \quad (5.25)$$

□

5.4 Solution Method

The solution method that we describe here is well suited for the class of problems (5.16)–(5.19) where U is quadratic. We leave the more general case for future research. Even with this simplification, we believe that the development of a solution technique for (5.16)–(5.19) is a valuable endeavor.

The mathematical program (5.16)–(5.19) is an instance of a Bilevel Programming problem. The first problem of this class, introduced by [146], modeled the interaction of two firms. The leader firm, which moves first, selects its production quantity knowing that the follower will observe its decision and respond accordingly. That is, in defining its strategy the leader takes into account the reaction of the follower, which in turn depends on the leader's decision. Bilevel optimization problems generalize this setting. A program in this class has an upper level (leader's problem) that has a set of constraints which are the solution set of subordinate (follower) problems.

When a solution set has more than one element, one can take two approaches. One, called pessimistic, assumes that followers do not cooperate with the leader; the other, often referred to as optimistic, assumes the opposite. Under this approach, the leader can select any element of the solution set of a follower. In this paper, we take the optimistic approach.

Because in our setting the subordinate problems (which (5.17) describes) are convex, their first order necessary conditions are also sufficient. Thus, we can use these conditions to express the solution sets of the subordinate problems analytically. In doing so we are effectively casting the Bilevel Problem (5.16)–(5.19) as a Mathematical Program with Equilibrium Constraints (MPEC).

5.4.1 Formulating the Ramsey Problem as an MPEC

Let ν^h be the Lagrange multiplier of the constraint of problem (5.17). For those customers under rate h , the Lagrangean of this problem is

$$L = E [U(\mathbf{d}^h; \boldsymbol{\theta}^h) + M_h - (\mathbf{d}^h)^\top \Lambda^h \mathbf{p}^h + (b^h - A^h \mathbf{d}^h)^\top \boldsymbol{\nu}^h], \quad (5.26)$$

and the first order necessary conditions are

$$\nabla U(d_\omega^h; \theta_\omega^h) - \Lambda^h p_\omega^h - (A^h)^\top \nu_\omega^h = 0, \quad (5.27)$$

$$0 \leq \nu_\omega^h \perp b^h - A^h d_\omega^h \geq 0 = 0 \quad (5.28)$$

for all $\omega \in \Omega$ and $h = 1, \dots, n$, with the symbol \perp indicating that the vectors must be orthogonal. We reformulate the Bilevel Model (5.16)–(5.19) replacing (5.17) with conditions (5.27)–(5.28). Henceforth we refer to this reformulation as RMPEC. The new problem falls within a class of non-linear programs which are particularly difficult to solve: MPEC. Because these problems fail to satisfy the Mangasarian-Fromovitz constraint qualifications at every feasible point, there are no convergence guarantees for standard non-linear methods. As a consequence, researchers have developed specialized algorithms to find stationary points for these programs [61]. While these algorithms may work well for small- to middle-sized instances, due to the combinatorial nature of MPECs [95], finding an exact solution for large-scale instances becomes impractical. Because the instances that we are interested to tackle are large-scale, in what follows we develop a specialized approximation method to find stationary points for large-scale instances of the RMPEC. Our technique is based upon the Alternating Direction Method of Multipliers (ADMM).

5.4.2 Decomposing the Problem

The key idea behind the ADMM algorithm is to combine the decomposability of the dual ascent or dual subgradient methods with the superior convergence of the method of multipliers. The former group of methods aims to find a solution to the original (primal) problem by solving its dual with an iterative procedure. This is often useful when the dual problem has a structure that permits simplifying its resolution. One important example is when the primal has coupling constraints—constraints that link a group of variables together. One can construct a dual relaxing these constraints. By doing so, the computation of a dual step can be decoupled and distributed, improving the scalability of the problem. A dual step is the multiplication of a step direction by a step size. In the case of the dual ascent method the step direction is the gradient of the dual function at the current point. Its computation requires the Lagrangian having a unique optimum at such a point. While this holds in many problems in some important cases, such as for mixed integer programs, the condition does not hold. A researcher can overcome this difficulty with a dual subgradient method. These use subgradients as steps directions. While consecutive iterations may not improve the objective function of the dual problem, provided that the step size is selected properly, the subgradient method reduces the distance between the current dual solution and the dual optimum at each iteration. One problem when using this method is that the computation of the step size is not straightforward. The correct magnitude depends on the optimal value, which is not known. As a result, the performance of this method is highly dependent on the specific structure of the problem at hand, being very effective in some cases and showing very slow convergence or an oscillating behavior in others.

A notable approach that overcomes the limitations of the dual ascent or subgradient methods is the method of multipliers. The crucial idea behind this technique is to strengthen the convexity of the objective function of the original problem to facilitate the search of an optimal solution. One way of achieving this is by adding a term to the objective which penalizes the violation of a set of

constraints. Since what is penalized are constraint violations, the original and the problem with the penalty—henceforth the augmented problem—are equivalent. However, under mild conditions the latter's structure permits the application of a dual ascent method, with good global convergence properties. The main drawback of this technique is that the penalty term makes it impossible to decouple and distribute the computation of the dual steps, which hinders the practical use of this approach for large-scale problems.

The alternating direction method of multipliers seeks to combine the best properties of the dual ascent (or subgradient) method and the method of multipliers. It does so by introducing a slight modification to the dual step computation of the associated augmented problem. For concreteness and because we will use it later, we now provide a description of the algorithm ADMM for a class of problems that, as we show in the next subsection, contains the RMPEC.

Let F and G_j be multivariate smooth functions, with $j \in \{0, 1, \dots, J\}$ and J possibly large. Let $Z = Z_0 \times Z_1 \times \dots \times Z_J$ be the domain of F , with Z_j a convex set contained in \mathbb{R}^{m_j} . The domain of each G_j is $Z_j \times \bar{\mathcal{W}}_j$, also a convex set, and we denote (z_j, \bar{w}_j) a generic element of this set. The range of F is $R_F \subseteq \mathbb{R}$ and that of G_j is $R_{G_j} \subseteq \mathbb{R}^{\iota_j}$, with ι_j some positive integer. The problem of interest is

$$\min_{(z, \bar{w}) \in Z \times \bar{\mathcal{W}}} \{F(z) : G_j(z_j, \bar{w}_j) \leq 0 \ \forall j = 1, \dots, J\}, \quad (5.29)$$

where $\bar{\mathcal{W}} = \bar{\mathcal{W}}_1 \times \dots \times \bar{\mathcal{W}}_J$.

Suppose we need to decouple this problem, for instance, because the constraints that the functions G_j define make the problem hard to solve. To this end, consider the equivalent program

$$\min F(z) \quad (5.30)$$

subject to

$$G_j(\hat{w}_j, \bar{w}_j) \leq 0 \ \forall j = 1, \dots, J, \quad (5.31)$$

$$z_j = \hat{w}_j \ \forall j = 1, \dots, J, \quad (5.32)$$

$$(z, \bar{w}) \in Z \times \bar{\mathcal{W}}, \quad (5.33)$$

We could use a dual subgradient method relaxing the constraints (5.32) to decouple this problem. But, as we discussed before, this approach can have poor performance. Instead, we use the ADMM algorithm. As the method of multipliers, this technique also defines an Augmented Lagrangean

$$L_\rho(w, z; \gamma) = F(z) + \sum_{j=1}^J \gamma_j^\top (z_j - \hat{w}_j) + \frac{\rho}{2} \|z_j - \hat{w}_j\|^2, \quad (5.34)$$

where $w = (\hat{w}_1^\top, \dots, \hat{w}_J^\top, \bar{w}_1^\top, \dots, \bar{w}_J^\top)^\top$, $\gamma = (\gamma_1^\top, \dots, \gamma_J^\top)^\top$ is the block vector of the dual variables of (5.32) and $\rho > 0$ a penalty. The key difference between ADMM and the method of multipliers is that the former do not minimizes at each iteration $L_\rho(w, z; \gamma)$. Instead, it decreases this function in two steps, a w - and z -minimization step. Before presenting this algorithm, we define the matrices B_1, B_2 such that (5.32) is equivalent to

$$B_1 w + B_2 z = 0. \quad (5.35)$$

Algorithm 5 describes ADMM applied to problem (5.29).

Algorithm 5 Alternating Direction Method of Multipliers

- 1: **Initialization:** Given $z^0 \in Z$, $\gamma^0 = 0$, two tolerances $e^{pri} > 0$ and $e^{dual} > 0$, and a primal and a dual residual r^0, s^0 , such that $\|r^0\| > e^{pri}$ and $\|s^0\| > e^{dual}$
 - 2: **while** $(\|r^k\| > e^{pri}) \wedge (\|s^k\| > e^{dual})$ **do**
 - 3: *w-minimization:* $w^{k+1} \leftarrow \arg \min_w \{L_\rho(w, z^k; \gamma^k) : G(\hat{w}_j, \bar{w}_j) \leq 0, \bar{w}_j \in \bar{\mathcal{W}}_j, \forall j\}$
 - 4: *z-minimization:* $z^{k+1} \leftarrow \arg \min_z \{L_\rho(w^{k+1}, z; \gamma^k) : z \in Z\}$
 - 5: $\gamma^{k+1} \leftarrow \gamma^k + \rho(B_1 w^{k+1} + B_2 z^{k+1})$
 - 6: $r^{k+1} \leftarrow B_1 w^{k+1} + B_2 z^{k+1}$
 - 7: $s^{k+1} \leftarrow \rho B_1^\top B_2 (z^{k+1} - z^k)$
 - 8: **end while**
-

The *w-minimization* step can be distributed noting that if z is fixed then the Augmented Lagrangian can be decoupled. Thus, another way of obtaining w^{k+1} is by assigning to

$$(\hat{w}_j^{k+1}, \bar{w}_j^{k+1}) \leftarrow \arg \min_{(\hat{w}_j, \bar{w}_j)} \left\{ -\gamma_j^k \cdot \hat{w}_j + \frac{\rho}{2} \|z_j^k - \hat{w}_j\|^2 : G(\hat{w}_j, \bar{w}_j) \leq 0, \bar{w}_j \in \bar{\mathcal{W}}_j \right\} \quad (5.36)$$

for every $j = 1, \dots, J$.

It is possible to show that when the functions F, G_j are convex (possibly nonsmooth) and if strong duality holds then the algorithm converges to a global optimum [16]. Even if F is nonconvex but smooth, it can still converge to stationary solutions [68]. When the constraints $G(\hat{w}_j, \bar{w}_j) \leq 0$ define nonconvex regions then one can use the algorithm as a heuristic. [138] show that for problems with quadratic objectives and nonconvex separable constraints, the algorithm can rapidly converge to approximated solutions, and in many cases to the global optimum.

A closely related method, extensively used in large-scale stochastic optimization, is Progressive Hedging (PH). As the alternating direction method of multipliers, this algorithm—introduced by [123]—is a specialization of the Proximal Point algorithm [123]. While PH converges to stationary points for a large class of stochastic optimization problems, for models involving discrete variables the method becomes a heuristic. In practice, it has proven to be a very effective technique to tackle large-scale, stochastic mixed integer programs (see, e.g., [93]; [92]), and more recently to solve stochastic MPECs (see, e.g., [51]; [63]).

We propose using ADMM as a heuristic for the particular stochastic MPEC that we introduce in this paper. The experience of [51], [63] and [138] suggests that this is a suitable approach, and Subsection 5.4.4 provides additional evidence.

5.4.3 Implementing ADMM

Before showing how we implement the alternating direction method of multipliers, we make some clarifications that ease the understanding of our approach. Henceforth we treat a random vector ξ also as a block vector, i.e., $\xi = (\xi_1^\top, \dots, \xi_{|\Omega|}^\top)^\top$ is an alternative representation. If ξ is a block

vector of random vectors, that is $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\top, \dots, \boldsymbol{\xi}_n^\top)^\top$, another way of expressing this object is $\boldsymbol{\xi} = (\xi_{11}^\top, \dots, \xi_{|\Omega|1}^\top, \dots, \xi_{1n}^\top, \dots, \xi_{|\Omega|n}^\top)^\top$.

We start our description showing that the problem RMPEC is an instance of (5.29). To see this, define $z := (\alpha^\top, \mathbf{d}^\top, \mathbf{p}^\top)^\top$, the constraint set $Z := \mathcal{A} \times \mathcal{D} \times \mathcal{P}$, where $\mathcal{D} = \mathbb{R}^{|\Omega|\kappa_1^1} \times \dots \times \mathbb{R}^{|\Omega|\kappa_n^1}$ and κ_h^1 is the number of columns of A_h . In addition, define the functions

$$F(z) := C \left(\sum_{h=1}^n \alpha_h \Psi_h \mathbf{d}^h \right) - \sum_{h=1}^n \alpha_h E [U(\mathbf{d}^h; \boldsymbol{\theta}^h)], \quad (5.37)$$

and

$$G_{\omega h}(z_{\omega h}, \bar{w}_{\omega h}) := \begin{bmatrix} \nabla U(d_\omega^h; \theta_\omega^h) - \Lambda^h p_\omega^h - (A^h)^\top \nu_\omega^h \\ -\nabla U(d_\omega^h; \theta_\omega^h) + \Lambda^h p_\omega^h + (A^h)^\top \nu_\omega^h \\ (b^h - A^h d_\omega^h)^\top \nu_\omega^h \\ -(b^h - A^h d_\omega^h)^\top \nu_\omega^h \\ -b^h + A^h d_\omega^h \end{bmatrix} \quad \forall \omega \in \Omega, \quad h = 1, \dots, n, \quad (5.38)$$

where $z_{\omega h} := (d_\omega^h, p_\omega^h)$, $\bar{w}_{\omega h} := \nu_\omega^h$, and $\bar{\mathcal{W}}_{\omega h} = \mathbb{R}_+^{\kappa_h^2}$, with κ_h^2 the number of rows of A_h .

In order to decouple the problem, we duplicate the vectors $z_{\omega h}$ introducing the variables $\hat{w}_{\omega h} = (\hat{d}_{\omega h}, \hat{p}_{\omega h})$ and the coupling constraints

$$\alpha_h(d_\omega^h - \hat{d}_\omega^h) = 0 \quad \forall \omega \in \Omega, \quad h = 1, \dots, n \quad \text{and} \quad (5.39)$$

$$p_\omega^h - \hat{p}_\omega^h = 0 \quad \forall \omega \in \Omega, \quad h = 1, \dots, n, \quad (5.40)$$

Because we will use it later, we now define $Q(w, z)$ as a block vector function that has in each block (ω, h) the left hand side of (5.39) and (5.40). In addition, we call SMPEC the stochastic MPEC that results from replacing (5.32) with (5.39) and (5.40) in (5.30)–(5.33). Note that, while (5.32) defines a set of linear constraints, those in (5.39) are nonlinear. This does not alter the structure of Algorithm 5, however, it does changes the update of s^{k+1} . We will describe this update in detail later in this subsection. In what follows, we concentrate in the optimization steps.

The Augmented Lagrangean of the reformulated problem is

$$\begin{aligned} L_\rho(\mathbf{w}, z; \boldsymbol{\gamma}) = & - \sum_{h=1}^n \alpha_h E \left[U(\mathbf{d}^h; \boldsymbol{\theta}^h) - (\mathbf{d}^h - \hat{\mathbf{d}}^h)^\top \boldsymbol{\gamma}_d^h - \frac{\rho}{2} \alpha_h \|\mathbf{d}^h - \hat{\mathbf{d}}^h\|^2 \right] \\ & + \sum_{h=1}^n E \left[(\mathbf{p}^h - \hat{\mathbf{p}}^h)^\top \boldsymbol{\gamma}_p^h + \frac{\rho}{2} \|\mathbf{p}^h - \hat{\mathbf{p}}^h\|^2 \right] + C \left(\sum_{h=1}^n \alpha_h \Psi_h \mathbf{d}^h \right), \end{aligned} \quad (5.41)$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_d^h, \boldsymbol{\gamma}_p^h)$ corresponds to the block vector of Lagrange multipliers of the constraints (5.39) and (5.40), and $\mathbf{w} := (\hat{\mathbf{d}}^\top, \hat{\mathbf{p}}^\top, \boldsymbol{\nu}^\top)^\top$.

The w -minimization step

In order to update w^{k+1} , we solve for every $\omega \in \Omega$ and $h \in \{1, \dots, H\}$ the MPEC that follows

$$\min_{(\hat{d}_{\omega h}, \hat{p}_{\omega h}, \nu_{\omega h})} \alpha_h \left[(d_{\omega}^h - \hat{d}_{\omega}^h)^{\top} \gamma_{\omega d}^h + \frac{\rho}{2} \alpha_h \|d_{\omega}^h - \hat{d}_{\omega}^h\|^2 \right] + (p_{\omega}^h - \hat{p}_{\omega}^h)^{\top} \gamma_{\omega p}^h + \frac{\rho}{2} \|p_{\omega}^h - \hat{p}_{\omega}^h\|^2 \quad (5.42)$$

$$\text{subject to (5.27) and (5.28),} \quad (5.43)$$

and assign its solution to $(\hat{d}_{\omega h}^{k+1}, \hat{p}_{\omega h}^{k+1}, \nu_{\omega h}^{k+1})$.

There are various approaches to solve this optimization problem. We refer the reader to [61] for a comprehensive review. Here we describe the approach we take, which casts the MPEC as a Mixed-Integer Quadratic Program (MIQP). The technique, proposed by [58], introduces integer variables and new constraints to reformulate the complementarity conditions. Letting $\sigma_{\omega}^h \in \{0, 1\}^{\kappa_h^2}$, $\bar{M}_{\omega}^h, \tilde{M}_{\omega}^h > 0$ scalars, and e a vector of ones of dimension κ_h^2 , in our setting the procedure involves replacing (5.28) with the following constraints

$$\bar{M}_{\omega}^h \sigma_{\omega}^h \geq b^h - A^h d_{\omega}^h, \quad (5.44)$$

$$\tilde{M}_{\omega}^h (e - \sigma_{\omega}^h) \geq \nu_{\omega}^h, \quad (5.45)$$

$$b^h - A^h d_{\omega}^h \geq 0, \quad (5.46)$$

$$\nu_{\omega}^h \geq 0, \quad (5.47)$$

$$\sigma_{\omega}^h \in \{0, 1\}^{\kappa_h^2}. \quad (5.48)$$

Researchers can solve the reformulated problem (5.42), (5.27), (5.44)–(5.48), which we refer to as w -MPEC, using standard MIQP techniques.

In principle, the approach of [58] can be inaccurate and inefficient. Inaccurate because if $\bar{M}_{\omega}^h, \tilde{M}_{\omega}^h$ are too small, $b^h - A^h d_{\omega}^h$ or ν_{ω}^h may be far from the optimum of (5.42)–(5.43). On the other hand, increasing these scalars too much can lead to inefficiencies because the problem could be ill-conditioned. Besides, even if they are selected properly, since MIQP is in NP-complete [116], it can take a long time for a solver to reach a solution of acceptable quality. In practice, for the class of problems that we focus on, non of these concerns posed significant difficulties to use this approach. Based on the physical nature of the constraints of the customer problem (5.46), one can find adequate values for \bar{M}_{ω}^h and, with some additional work, for \tilde{M}_{ω}^h ; and while the complexity of the problem translated into slow performance, we were able to overcome this issue by providing the MIQP solver with good starting points.

A good starting point decreases the solution time of w -MPEC by pruning several suboptimal branches of the Branch and Bound tree, and by providing an adequate range for \tilde{M}_{ω}^h . The procedure we used to construct a good starting point exploits that, in general, solutions across iterations do not differ as much, and that depending on the magnitude of α_h the first two terms of (5.42) might be irrelevant. Then, one way of constructing a starting point is using the price of the previous iteration and solve the customer problem to find the associated demand and dual variables. In addition, assuming the first two terms of (5.42) are irrelevant (because α_h is relatively small) then, one can construct another starting point finding $(\hat{p}_{\omega}^h)^* = \arg \min_{\hat{p}_{\omega}^h} \{ (p_{\omega}^h - \hat{p}_{\omega}^h)^{\top} \gamma_{\omega p}^h + \frac{\rho}{2} \|p_{\omega}^h - \hat{p}_{\omega}^h\|^2 \}$

and solving the customer problem for this price. A third technique is to simply use \hat{p}_ω^h as input to the customer problem. Then, the starting point is the one with the smallest objective value for w -MPEC.

The z -minimization step

Note that the z -minimization problem can be separated into two subproblems. One that finds the optimal (α, \mathbf{d}) and another that optimizes \mathbf{p} . This is possible because there are no constraints coupling these block of variables. The problem that optimizes \mathbf{p} is

$$\min_{\mathbf{p}} \left\{ \sum_{h=1}^n E \left[(\mathbf{p}^h - \hat{\mathbf{p}}^h)^\top \gamma_p^h + \frac{\rho}{2} \|\mathbf{p}^h - \hat{\mathbf{p}}^h\|^2 \right] : \mathbf{p} \in \mathcal{P} \right\}, \quad (5.49)$$

which depending on the structure of \mathcal{P} (a convex set), can be solved either analytically or using a standard convex optimization solver. To find the optimal (α, \mathbf{d}) , we solve

$$\min_{(\alpha, \mathbf{d})} \left\{ C \left(\sum_{h=1}^n \alpha_h \Psi_h \mathbf{d}^h \right) - \sum_{h=1}^n \alpha_h E \left[U(\mathbf{d}^h; \boldsymbol{\theta}^h) - (\mathbf{d}^h - \hat{\mathbf{d}}^h)^\top \gamma_d^h - \frac{\rho}{2} \alpha_h \|\mathbf{d}^h - \hat{\mathbf{d}}^h\|^2 \right] : \alpha \in \mathcal{A} \right\}. \quad (5.50)$$

We could use a standard nonlinear solver to handle this problem. However, specially for problems of large size, this approach is less attractive than using specialized convex optimization algorithms. We are able to take this approach by casting the problem (5.50) as a conic program. In order to do so, we first rewrite the term inside the expectation in the objective function as follows

$$U(\hat{\mathbf{d}}^h; \boldsymbol{\theta}^h) + (\mathbf{d}^h - \hat{\mathbf{d}}^h)^\top (\nabla U(\hat{\mathbf{d}}^h; \boldsymbol{\theta}^h) - \gamma_d^h) + \frac{1}{2} (\mathbf{d}^h - \hat{\mathbf{d}}^h)^\top \left[\nabla^2 U(\hat{\mathbf{d}}^h; \boldsymbol{\theta}^h) - \rho \alpha_h I_h \right] (\mathbf{d}^h - \hat{\mathbf{d}}^h), \quad (5.51)$$

where we replaced $U(\mathbf{d}^h; \boldsymbol{\theta}^h)$ by its taylor expansion about $\hat{\mathbf{d}}^h$, and I_h is an identity of dimension κ_h^1 . Next, we introduce variables $\tilde{\mathbf{d}}^h = \alpha_h \mathbf{d}^h$, \mathbf{v}_0^h , \mathbf{v}_1^h , and write the following reformulation of the problem

$$\min_{(\alpha, \tilde{\mathbf{d}})} C \left(\sum_{h=1}^n \Psi_h \tilde{\mathbf{d}}^h \right) - \sum_{h=1}^n E \left[\alpha_h U(\hat{\mathbf{d}}^h; \boldsymbol{\theta}^h) + (\tilde{\mathbf{d}}^h - \alpha_h \hat{\mathbf{d}}^h)^\top (\nabla U(\hat{\mathbf{d}}^h; \boldsymbol{\theta}^h) - \gamma_d^h) - \mathbf{v}_0^h - \mathbf{v}_1^h \right] \quad (5.52)$$

subject to

$$2\alpha_h \mathbf{v}_0^h \geq (\tilde{\mathbf{d}}^h - \alpha_h \hat{\mathbf{d}}^h)^\top \left[-\nabla^2 U(\hat{\mathbf{d}}^h; \boldsymbol{\theta}^h) \right] (\tilde{\mathbf{d}}^h - \alpha_h \hat{\mathbf{d}}^h), \quad \forall h = 1, \dots, n \quad (5.53)$$

$$2\mathbf{v}_1^h \geq \rho (\tilde{\mathbf{d}}^h - \alpha_h \hat{\mathbf{d}}^h)^\top (\tilde{\mathbf{d}}^h - \alpha_h \hat{\mathbf{d}}^h), \quad \forall h = 1, \dots, n \quad (5.54)$$

$$\alpha \in \mathcal{A}. \quad (5.55)$$

To obtain a conic program, we introduce the variables ϕ_0^h , ϕ_1^h , \mathbf{v}_2^h , replace (5.53) with the condi-

tions

$$2\alpha_h \mathbf{v}_0^h \geq (\phi_0^h)^\top \phi_0^h, \quad \forall h = 1, \dots, n \quad (5.56)$$

$$\phi_0^h = \left[-\nabla^2 U(\hat{\mathbf{d}}^h; \boldsymbol{\theta}^h) \right]^{\frac{1}{2}} (\tilde{\mathbf{d}}^h - \alpha_h \hat{\mathbf{d}}^h), \quad \forall h = 1, \dots, n, \quad (5.57)$$

and (5.54) with

$$2\mathbf{v}_1^h \mathbf{v}_2^h \geq (\phi_1^h)^\top \phi_1^h, \quad \forall h = 1, \dots, n \quad (5.58)$$

$$\phi_1^h = \rho^{\frac{1}{2}} (\tilde{\mathbf{d}}^h - \alpha_h \hat{\mathbf{d}}^h), \quad \forall h = 1, \dots, n, \quad (5.59)$$

$$\mathbf{v}_2^h = 1, \quad \forall h = 1, \dots, n. \quad (5.60)$$

While constraints (5.56) and (5.58) define two rotated quadratic cones, the others are simply linear constraints. The problem that (5.52), (5.56)–(5.60) and (5.55) define is a conic program. Researchers can solve it using an off-the-shelf conic optimization package.

Updates

After the two optimization steps Algorithm 5 updates the dual variable γ^{k+1} , and residuals r^{k+1} and s^{k+1} . We now show how we adapt these steps for SMPEC. The dual variable and (primal) residual r^{k+1} update in a similar fashion. That is,

$$\gamma^{k+1} \leftarrow \gamma^{k+1} + \rho Q(w^{k+1}, z^{k+1}), \quad (5.61)$$

$$r^{k+1} \leftarrow Q(w^{k+1}, z^{k+1}). \quad (5.62)$$

However, the update of the (dual) residual s^{k+1} is different. In order to derive an update rule for this vector, we follow a line of reasoning similar to that used in the standard ADMM algorithm. In this technique, if the norms of r^{k+1} and s^{k+1} are small then, provided that the problem is convex, the algorithm converged to an optimal solution [16]. We now discuss why this is the case considering (5.30)–(5.33), and then show how we can adapt this reasoning to construct a residual update for SMPEC.

Let \mathcal{X} be a convex set. We denote $\mathbb{I}_{\mathcal{X}}(x)$ the indicator function, which is 0 when $x \in \mathcal{X}$ and ∞ otherwise. In addition, define the set $\mathcal{W} := \{(\hat{w}, \bar{w}) : G_j(\hat{w}_j, \bar{w}_j) \leq 0, \forall j, \bar{w} \in \bar{\mathcal{W}}\}$. The Lagrangian of (5.30)–(5.33) is

$$L(w, z; \mu, \gamma) = F(z) + \gamma^\top (B_1 w + B_2 z) + \mathbb{I}_{\mathcal{W}}(w) + \mathbb{I}_Z(z). \quad (5.63)$$

Assuming that the Slater constraint qualification holds then, the necessary and sufficient condition for (w^*, z^*) to be optimal is $0 \in \partial_{(z,w)} L(w^*, z^*; \mu, \gamma)$, where $\partial_{(z,w)}$ denotes the subdifferential with respect to (z, w) [129]. Using subdifferential calculus rules (e.g., [124]), we can write this condition as follows

$$0 \in B_1^\top \gamma + N_{\mathcal{W}}(w), \quad (5.64)$$

$$0 \in \nabla F(z) + B_2^\top \gamma + N_Z(z), \quad (5.65)$$

with $N_{\mathcal{X}}(x)$ denoting the normal cone of \mathcal{X} at x .

On the other hand, we have that the z -minimization step of Algorithm 5 is such that z^{k+1} satisfies

$$0 \in \partial_{\omega} L_{\rho}(w^{k+1}, z^{k+1}; \gamma^k) + N_Z(z^{k+1}) \quad (5.66)$$

$$= \nabla F(z^{k+1}) + B_2^{\top} \gamma^k + \rho B_2^{\top} r^{k+1} + N_Z(z^{k+1}) \quad (5.67)$$

$$= \nabla F(z^{k+1}) + B_2^{\top} \gamma^{k+1} + N_Z(z^{k+1}), \quad (5.68)$$

where the last equality follows from the update rule for γ . In other words, the points that Algorithm 5 generates always satisfy (5.65). The same does not hold for condition (5.64). Indeed, w^{k+1} satisfies

$$0 \in \partial_{\omega} L_{\rho}(w^{k+1}, z^k; \gamma^k) + N_{\mathcal{W}}(w^{k+1}) \quad (5.69)$$

$$= B_1^{\top} \gamma^k + \rho B_1^{\top} (B_1 w^{k+1} + B_2 z^k) + N_{\mathcal{W}}(w^{k+1}) \quad (5.70)$$

$$= B_1^{\top} \gamma^k + \rho B_1^{\top} r^{k+1} + N_{\mathcal{W}}(w^{k+1}) + \rho B_1^{\top} B_2 (z^k - z^{k+1}), \quad (5.71)$$

$$= B_2^{\top} \gamma^{k+1} + N_{\mathcal{W}}(w^{k+1}) + s^{k+1}, \quad (5.72)$$

which differs from (5.64). However, if s^{k+1} vanishes at some iteration $k+1$ then, (5.72) becomes (5.64). If in addition, $r^{k+1} = 0$ then (w^{k+1}, z^{k+1}) is optimal for (5.30)–(5.33).

We can follow a similar strategy to define the update of s^k in the context of SMPEC. However, we need to address two properties that distinguish this problem from (5.30)–(5.33). One is that in our setting the coupling constraints (5.39) and (5.40) are nonlinear, the other that $G_{\omega h}(\hat{w}_{\omega h}, \bar{w}_{\omega h})$ are nonconvex functions. Because of these, we cannot use techniques from convex analysis in order to derive an update rule for s^k . However, if F , $G_{\omega h}$ and Q were locally Lipchitz-continuous,³ and the sets Z and \mathcal{W} were closed, we could leverage classic results from nonsmooth analysis. It easy to see that Z and \mathcal{W} are closed sets. We now show that F , $G_{\omega h}$ and Q are locally Lipchitz-continuous functions.

First, note that convex or smooth functions are locally Lipchitz-continuous [35], and that the summation and composition preserve this property. The function $F(z)$, defined in (5.37), is locally Lipchitz-continuous because it is the summation of two locally Lipchitz-continuous functions. The first term in the right hand side of (5.37) has this property since it is the composition of a convex with a smooth function, both locally Lipchitz-continuous; the second term is a smooth function of (α, \mathbf{d}) , and thus locally Lipchitz-continuous. In addition, $G_{\omega h}(\hat{w}_{\omega h}, \bar{w}_{\omega h})$ are smooth vector functions since all their components are smooth, which is also the case for $Q(w, z)$.

To write in a simple manner necessary conditions for optimality, we first introduce the block diagonal matrices Π_d and Π_p such that the block (ω, h) is equal to $q_{\omega} I_{\kappa_h^1}$ for the former and $q_{\omega} I_T$ for the latter, with q_{ω} the probability that ω occurs. In addition, we define the two-block diagonal matrix Π , having in its first block Π_d and Π_p in the second. It is direct to verify that for any pair of

³For a definition of local Lipchitz-continuity, we refer the reader to [35].

vectors, $(\mathbf{d}', \mathbf{p}')$ and $(\mathbf{d}'', \mathbf{p}'')$, with the dimensions of (\mathbf{d}, \mathbf{p}) , we have

$$(\mathbf{d}', \mathbf{p}')^\top \Pi(\mathbf{d}'', \mathbf{p}'') = E \left[\sum_{h=1}^n (\mathbf{d}'_h)^\top \mathbf{d}''_h + (\mathbf{p}'_h)^\top \mathbf{p}''_h \right] \quad (5.73)$$

We now define the following Lagrangian for the problem SMPEC

$$L(w, z; \mu, \gamma, \eta) = \mu F(z) + \gamma^\top \Pi Q(w, z) + \eta \|(\mu_1, \mu_2)\| d_{\mathcal{W} \times \mathcal{Z}}(w, z), \quad (5.74)$$

where η is some positive scalar and $d_{\mathcal{X}}(x) = \inf \{\|x - x'\| : x' \in \mathcal{X}\}$. Then, in virtue of Theorem 6.1.1 in [35], also known as the Lagrange multiplier rule, we have that if (w^*, z^*) solves globally or locally SMPEC

$$\exists \mu \geq 0 \text{ and } \gamma \text{ not all zero, such that } 0 \in \partial_{(w,z)} L(w^*, z^*; \mu, \gamma, \eta), \quad (5.75)$$

for every $\eta \geq \hat{\eta}$, with $\hat{\eta}$ the Lipchitz constant of the function $[F, Q]$ in a neighborhood of (w^*, z^*) , and $\partial_{(w,z)}$ denoting the Clarke subdifferential (see Definition 7.3.4 in [129]). Using calculus rules for this type of subdifferentials (see, e.g., [35]), we can rewrite the expression $0 \in \partial_{(w,z)} L(w, z; \mu, \gamma, k)$ as follows

$$0 \in \mu \nabla F(z^*) + J_z Q(w^*, z^*)^\top \Pi \gamma + N_Z(z^*) \quad (5.76)$$

$$0 \in J_w Q(w^*, z^*)^\top \Pi \gamma + N_{\mathcal{W}}(w^*), \quad (5.77)$$

with J the standard Jacobian, and $N_{\mathcal{X}}(x)$ denoting the normal cone of \mathcal{X} at x , defined in [35].

As in the convex case, we now analyze what first order conditions w^{k+1} and z^{k+1} satisfy. Before, we note that one can write the augmented Lagrangean of SMPEC, defined in (5.41), as follows

$$L_\rho(w, z; \gamma) = F(z) + \gamma^\top \Pi Q(w, z) + \frac{\rho}{2} Q(w, z)^\top \Pi Q(w, z). \quad (5.78)$$

As we showed before, the problem of the z -minimization step is convex. Thus, we have z^{k+1} satisfies

$$0 \in \nabla F(z^{k+1}) + J_z Q(w^{k+1}, z^{k+1})^\top \Pi \gamma^k + \rho J_z Q(w^{k+1}, z^{k+1})^\top \Pi Q(w^{k+1}, z^{k+1}) + N_Z(z^{k+1}) \quad (5.79)$$

$$= \nabla F(z^{k+1}) + J_z Q(w^{k+1}, z^{k+1})^\top \Pi \gamma^{k+1} + N_Z(z^{k+1}), \quad (5.80)$$

which is exactly (5.76), with $\mu = 1$. Before, analyzing the case of w^{k+1} , we introduce the following notation $Q_{k,k} := Q(w^k, z^k)$. In this case, using the Lagrange multiplier rule, we have that w^{k+1} satisfies

$$0 \in J_w Q_{k+1,k}^\top \Pi \gamma^k + \rho J_w Q_{k+1,k}^\top \Pi Q_{k+1,k} + N_{\mathcal{W}}(w^{k+1}) \quad (5.81)$$

$$= J_z Q_{k+1,k+1}^\top \Pi \gamma^{k+1} + N_{\mathcal{W}}(w^{k+1}) + s^{k+1}, \quad (5.82)$$

with

$$s^{k+1} = J_w Q_{k+1,k}^\top \Pi[\gamma^k + \rho Q_{k+1,k}] - J_w Q_{k+1,k+1}^\top \Pi \gamma^{k+1}. \quad (5.83)$$

In (5.81) we use that the Linear Independence constraint qualification for MPECs (MPEC LICQ) holds for the problem of the w -minimization step. In virtue of Theorem 2 in [128], and the fact that MPEC LICQ implies the MPEC Mangasarian-Fromovitz constraint qualification,⁴ we have that the multiplier of the objective function of the w -minimization problem is equal to 1.

The definition of dual residual in (5.83) is such that that our variation of the ADMM algorithm applied to the SMPEC stops at stationary point for this problem. While we do not prove convergence, Subsection 5.4.4 shows that the algorithm converges in most instances.

To gain a some intuition on when this residual becomes zero, we consider the case when α does not change between consecutive iterations. Before doing so, we note that

$$Q(w, z) = H(\alpha) \begin{bmatrix} \mathbf{d} - \tilde{\mathbf{d}} \\ \mathbf{p} - \hat{\mathbf{p}} \end{bmatrix} = H(\alpha) \Delta, \quad (5.84)$$

where $H(\alpha)$ is a block diagonal matrix such that the block corresponding to \mathbf{d}_h is equal to $\alpha_h I_{k_h}^{-1}$, and that corresponding to \mathbf{p} is an identity matrix. In addition, we observe that (5.83) can be expressed as follows

$$s^{k+1} = J_w Q_{k+1,k}^\top \Pi[\gamma^{k+1} + \rho(Q_{k+1,k} - Q_{k+1,k+1})] - J_w Q_{k+1,k+1}^\top \Pi \gamma^{k+1} \quad (5.85)$$

$$= [J_w Q_{k+1,k} - J_w Q_{k+1,k+1}]^\top \Pi \gamma^{k+1} + \rho J_w Q_{k+1,k}^\top \Pi [Q_{k+1,k} - Q_{k+1,k+1}], \quad (5.86)$$

which in view of (5.84) is equal to

$$H(\alpha^{k+1} - \alpha^k) \Pi \gamma^{k+1} - \rho H(\alpha^k) \Pi [H(\alpha^k) \Delta_{k+1,k} - H(\alpha^{k+1}) \Delta_{k+1,k+1}]. \quad (5.87)$$

If α does not vary between k and $k+1$

$$s^{k+1} = -\rho H(\alpha^k) \Pi H(\alpha^k) \begin{bmatrix} \mathbf{d}^k - \mathbf{d}^{k+1} \\ \mathbf{p}^k - \mathbf{p}^{k+1} \end{bmatrix} \quad (5.88)$$

$$= -\rho H(\alpha^k) \Pi \begin{bmatrix} \tilde{\mathbf{d}}^k - \tilde{\mathbf{d}}^{k+1} \\ \mathbf{p}^k - \mathbf{p}^{k+1} \end{bmatrix} \quad (5.89)$$

$$= -\rho E \left[\sum_{h=1}^n \alpha_h^k (\tilde{\mathbf{d}}_h^k - \tilde{\mathbf{d}}_h^{k+1}) + (\mathbf{p}^k - \mathbf{p}^{k+1}) \right], \quad (5.90)$$

a residual akin to that of Algorithm 5.

5.4.4 Testing the Performance of the Approach

We now focus on the computational performance of the method. In order to explore this, we work with moderately sized instances which we describe in detail in the next section. Here we

⁴See [155] for various definitions of constraint qualifications, including MPEC LICQ and MPEC Mangasarian-Fromovitz CQ.

concentrate on the convergence properties and the resolution times of the algorithm we propose for various sizes and parameter configurations. We change the size of the instances by varying the number of scenarios (or states of nature); the number of variables and complementarity constraints are affine transformations of this parameter. We consider 5 instance sizes, corresponding to 2, 3, 10, 25 and 50 scenarios, 2462, 3686, 12 254, 30 614 and 61 214 variables, or 1200, 1800, 6000, 15 000 and 30 000 complementarity constraints. For a given instance size, we run the algorithm for 40 different parameter configurations which corresponds to combinations of different rate structures, network constraints, and costs for the intermittent and distributed technologies.

We ran our experiment in an ASUS M51AC PC, with a 4 core processor Intel i7-4770 of 3.40 GHz, 16 GB of RAM, and a 64-bit Windows operating system. Of a total of 200 instances, we considered convergent the instances that meet the stopping criterion before 400 iterations. These corresponded to 188 or 94% of the total. The results we report in this subsection are those associated to the convergent instances. For convergence, we required the primal residual to be lower than or equal to 0.25% and a dual residual being lower than or equal to 2.5%. Intuitively, this criterion places more emphasis on the feasibility than the stationarity of the solution.

Before discussing the results, we note that in order to have a benchmark we attempted to run this experiment with an alternative algorithm. Specifically, we tried solving RMPEC with Knitro, a nonlinear optimization package which has specialized routines to solve MPECs (see [21]). We could not obtain results within 24 hours, even for the smallest instances we considered in this test. In contrast, with the same machine and using the technique we present in this paper, we were able to obtain results within a few hours for the majority of the instances.

Figure 5.1 shows the evolution of the average primal and dual residuals across iterations, for the instance sizes we tested. For most instances, after a few tens of iterations the algorithm finds a solution of reasonable quality, with small primal and dual residuals. The iterations that follow, while improving the solution quality, do this in a relatively slow fashion when compared with the improvement achieved in the first 50 iterations. Interestingly, this convergence behavior seems not to depend on the size, at least for the instances with 10 or more scenarios. We see a different behavior for the instances with 2 and 3 scenarios. While all instances of these sizes converged, the convergence metrics oscillate much more across iterations than the other instances. We only observe this oscillating behavior in the resolution of instances where the transmission is constrained.

In order to have an idea on the potential scalability of the algorithm, we plot in the Panel A of Figure 5.2 the resolution and average iteration times, and number of iterations for different instances sizes. We note that while the graph shows that there is an increase in total resolution times, the median is always below 2000 seconds. In fact, for more than 75% of the instances that converged, the algorithm found a stationary point in less than one hour. This result stands in sharp contrast to the more than 24 hours it took to the off-the-shelf solver to find a solution for instances with just 2 scenarios.

In addition, Panel A shows that the main driver of the rise in resolution times as the size of the instances increases is the average iteration time. As the number of states of nature rises, the distribution of the average time per iteration shifts upwards. This does not happen for the number of iterations. While its variance increases for instances with larger sizes, its median is always below 75. This suggests that if the time per iteration does not grows exponentially with the size of the

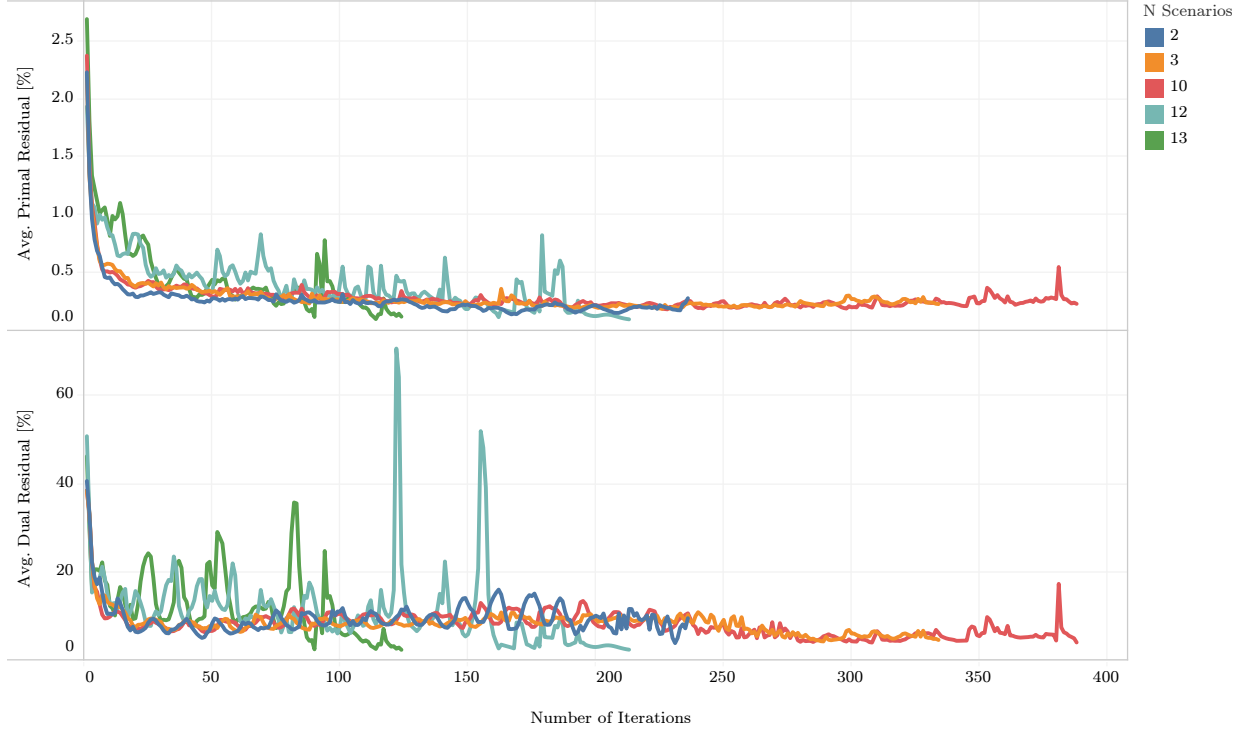


Figure 5.1: Primal and Dual Residual Evolution Across Iterations

problem, then it is unlikely that the total resolution time will grow at an exponential rate.

The plots in Panel B provide additional evidence supporting that this is the case. These plots graph the logarithm of the number scenarios versus the logarithm of the average iteration times and the number of iterations. The upper plot shows that a line of slope 0.94 and intercept -0.3 fits well the data. That is, for every additional scenario the average time per iteration increases on average by about a half of a second. On the other hand, the number of iterations increases in proportion to the square root of the number of scenarios.

These results highlight the potential of the algorithm to scale. The increase in the average time per iteration can be handled with more computing nodes, keeping approximately constant the duration of the w -minimization step. While the resolution time of the z -minimization step increases, it does so at a slower rate than linear.

5.5 A Simple Application

We conduct an analysis of rate structures using the method we develop in this paper. Our objective here is to show with an exercise the type of analysis a researcher can conduct using the technique. To keep the analysis simple, we perform this exercise with small sized instances, all with 3 representative scenarios. We explain bellow how we select this scenarios.

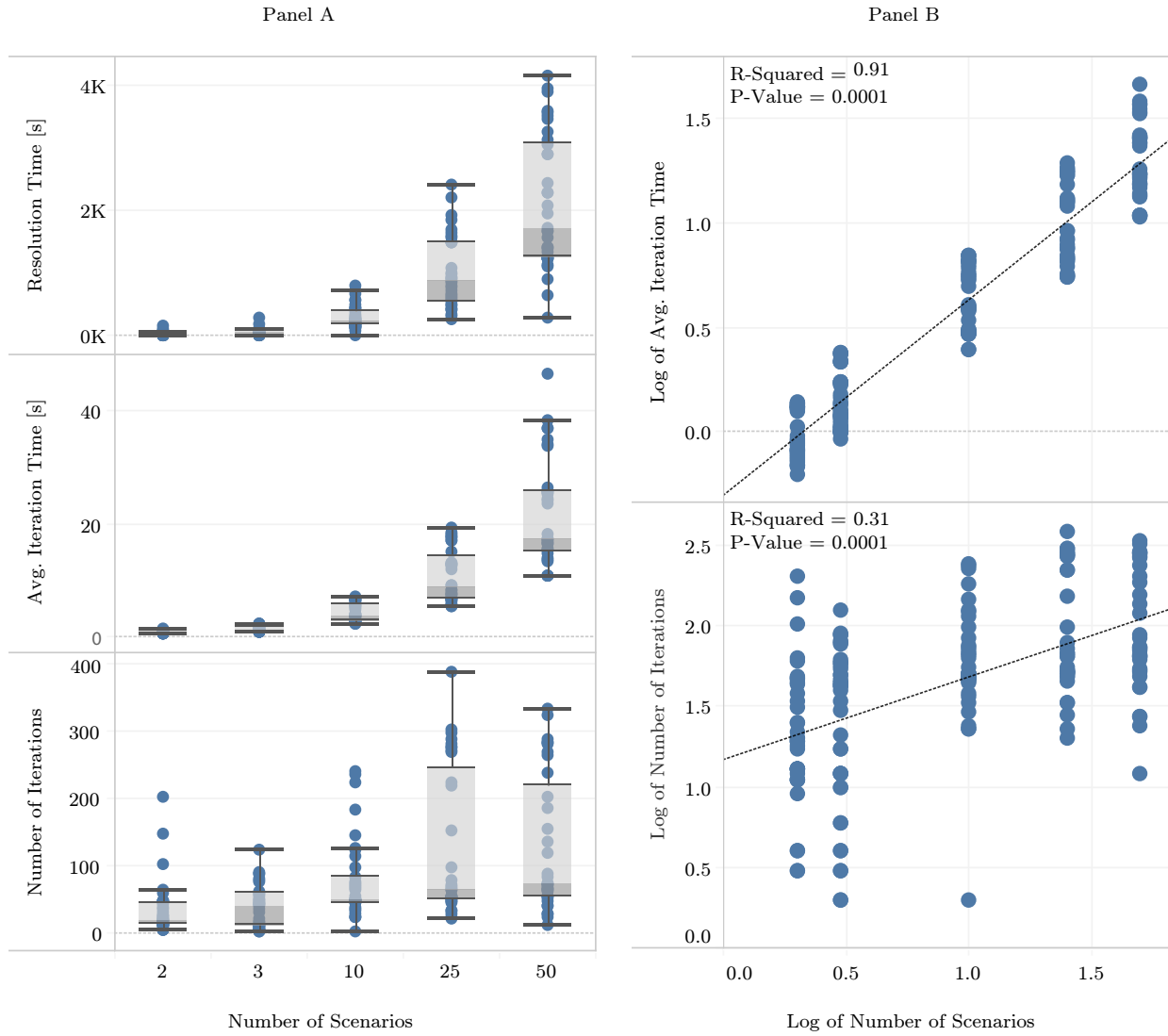


Figure 5.2: Iteration Metrics Versus Instances Sizes

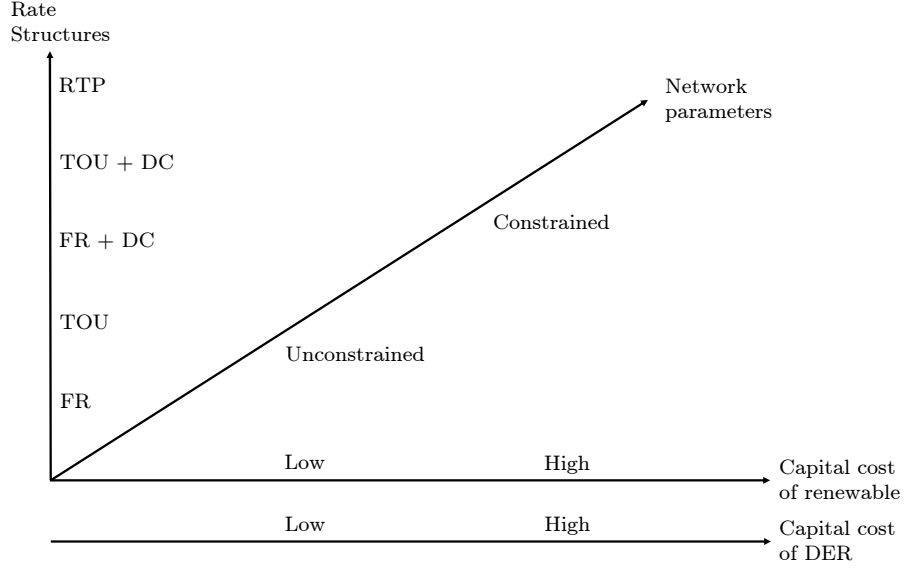


Figure 5.3: Structure of Analysis. An Instance Corresponds to the Combination of a Rate Structure, Network Parameter and the Renewable and DER Costs. There Are 30 Instances in Total

5.5.1 Designing the Analysis

In this analysis, we compare rate structures that have been proposed as possible alternatives for future electricity systems. We explore five tariffs, ranging from the simplest one, a flat-rate (FR), to the most complex, a real-time price (RTP). In addition, we study how the results change when changing the parameters of the transmission system, and the economics of the renewable generating technologies and distributed energy resources. Figure 5.3 describes the structure of the analysis.

As Figure 5.4 depicts, the network we consider has five buses, two load and three generation buses. In the base case scenario all branches have unlimited capacity; we change only the capacity of branch 1–4 to introduce congestion. In this network, there are five different technologies distributed as depicted in Figure 5.4. The economic parameters of the generation technologies are standard for capacity expansion studies (see, for instance, [40]). Table 5.1 summarizes these parameters.

Table 5.1: Economic Parameters of Supply-Side Technologies

		Base-load	Mid-merit	Peak	High-peak	Wind
Capital cost		207	85	27	16	225
Fixed O&M	k\$/MW-yr	69	21	16	11	40
Total fixed		227	106	43	27	265
Fuel		11	27	43	66	0
Variable O&M	\$/MWh	5	11	11	11	0
Total variable		16	38	54	77	0

Besides the network includes two load buses—1 and 2 in Figure 5.4, where a set of representative

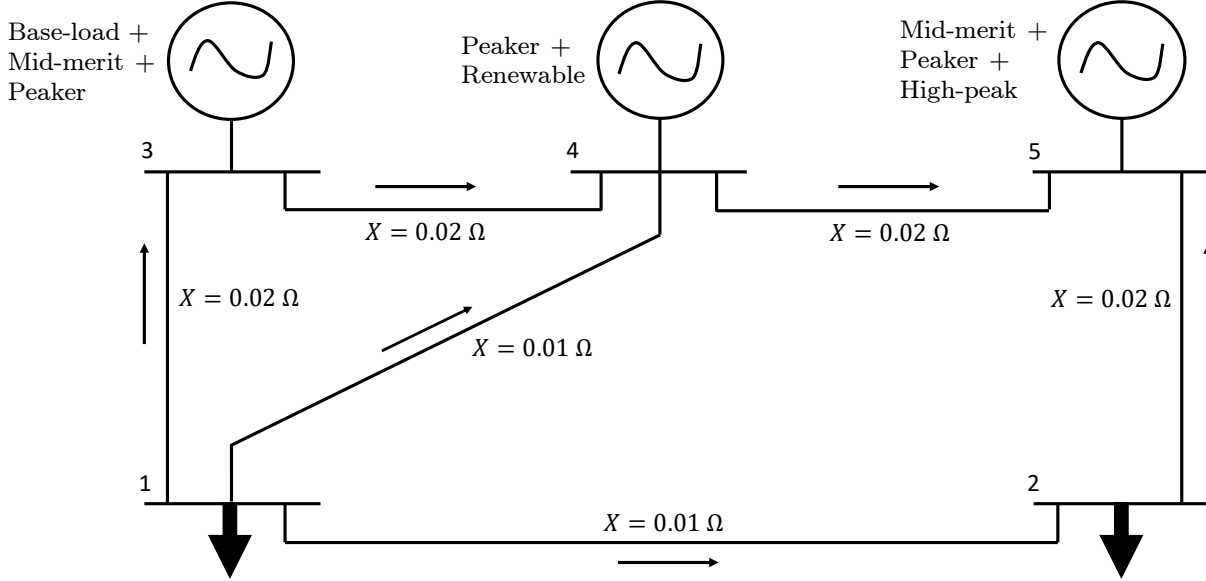


Figure 5.4: Network Model. The Letter X Denotes the Reactance of the Branch, and the Arrow the Default Direction of the Flows. Buses 1 and 2 Correspond to Loads, and 3, 4 and 5 to Generating Technologies

household are located. These households can differ in terms of their appliances and the DER they can adopt (for instance, because of differences in income). For this exercise, we consider two representative households per location. One can adopt DER, the other cannot. In addition, households located at bus 1 do not have air conditioning while those at bus 2 do. Figure 5.5 summarizes this configuration. As in Example 5.3.2, we model the consumption preferences for each household, assuming the utility function is additively separable. For the purposes of this exercise, we refer to the demand of all other appliances as baseline.

Another relevant input are the time series we use in this exercise. We model weather and consumption patterns with six time series, including temperature profiles and rooftop solar availability factors at each bus. We also include the baseline consumption of the households—which we assume is the same for both buses, and the availability factors of the wind power generator. All time series correspond to one year of hourly data, and we define a scenario as one day, or a 24 hour period. We sample from the data three representative scenarios using the importance sampling technique described in [110], which minimizes the distortions introduced by the sampling. Figure 5.6 depicts the time series we use and the scenarios we selected.

To the best of our knowledge, an analysis of these characteristics has not been undertaken before. The paper of [60] presents the closest attempt. The authors develop a model that allows them to estimate the parameters of a demand system with own and cross-price elasticities, and conduct a welfare comparison of time-varying rates and combinations with demand charges. The key difference between the analysis of [60] and the present is the representation of the supply side. These authors consider a simple supply cost function, which does not permit including renewable generation nor it allows modeling a network, or distributed energy resources. As we see below, different configurations of these factors can drastically impact the results.

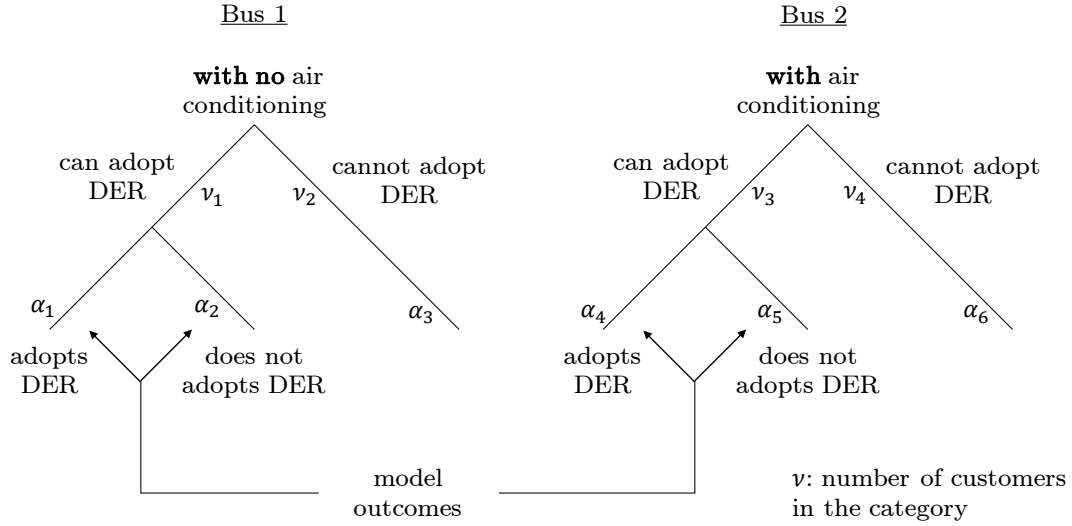


Figure 5.5: Household Configuration per Bus. There Are in Total 6 Different Households Types. The Final Number per Type, α , Is an Outcome of the Model, and is Constrained by the Number of Representative Households, ν .

5.5.2 Results

Welfare Analysis

We first analyze welfare differences with respect to the theoretically inferior rate, the flat rate. Table 5.2 shows these differences for all the parameter combinations and rates. We see that across all these configurations RTP is the structure that improves welfare the most; on the opposite extreme is the flat rate supplemented with a demand charge. The other two tariffs are close in the middle. On average, across all parameters configurations, the time of use and the combination of this structure with a demand charge obtain close to 90% of the welfare gains that RTP achieves, while the other structure only approximately 50%. These results suggests that while adding a demand charge to a flat rate structure can improve welfare (on average 1.93%), adding a demand charge does not have an effect when complementing a TOU structure. In addition, we see that the TOU outperforms the FR + DC structure, producing almost the same welfare improving effects as the RTP. We observe, however, that the relatively small differences between TOU and RTP are due to our definition of time-windows for the former tariff. This structure has 24 time-windows, 1 per hour; the main difference with respect to the RTP rate is that in the case of the TOU the price for a given hour cannot vary across scenarios. As the number of representative scenarios increases, we should observe a widening in the welfare gains that these two structures can achieve.

Table 5.2 shows, in addition, the affect of the other factors we explore in this exercise in the rate comparison. The network status has the most pronounced effect. When the network is constrained the welfare gains of switching from FR to any other rate are much higher, on average 31 times the gains that occur when the network is unconstrained. A model that does not account for a network implicitly assumes that there are no network constraints. This simple example shows that this assumption can have meaningful impacts on the rate comparison. A policy maker could conclude that given that the economic benefits of switching to a more complex rate are small, it is better to

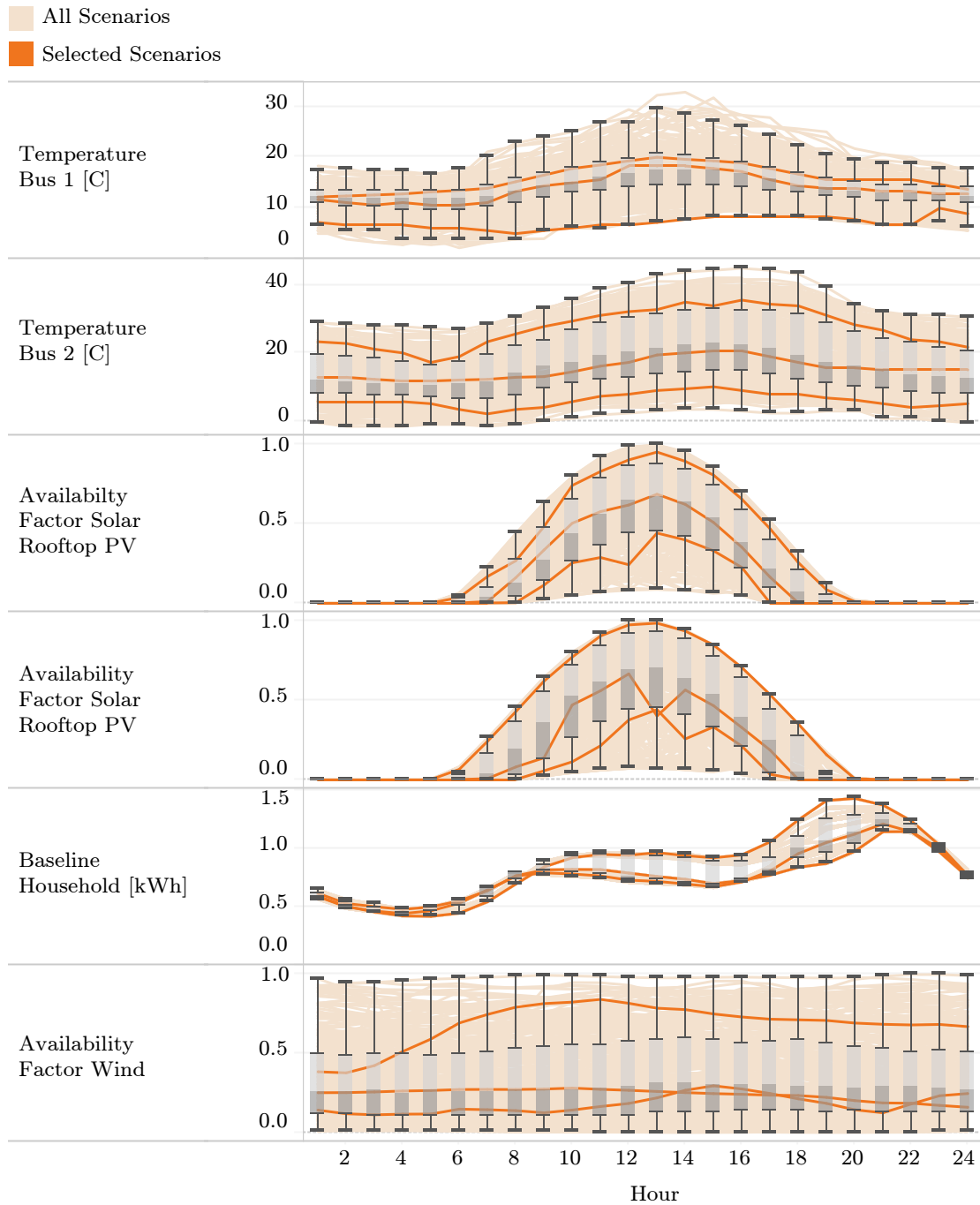


Figure 5.6: Time Series.

Table 5.2: Welfare Gains with Respect to FR in Percentages

Network Status ¹	Cost DER	Cost Renewable	Tariff			
			RTP	TOU	TOU + DC	FR + DC
const.	low	low	8.43	8.23	8.22	4.39
		high	8.01	7.81	7.71	4.70
	high	low	7.52	6.76	6.18	3.07
		high	7.52	6.78	6.49	3.08
unconst.	low	low	0.63	0.47	0.40	0.03
		high	0.63	0.41	0.40	-0.05
	high	low	0.49	0.32	0.30	0.11
		high	0.48	0.32	0.31	0.12

¹ const. = constrained, unconst. = unconstrained.

leave households enrolled in a FR tariff. The same conclusion would be hardly sustainable in the presence of the results of the constrained case.

The other factor that has an effect—though smaller—is the cost of the DER. When this technology is less expensive, the gain of switching from FR to a more complex rate is greater than when it has a higher cost. This happens because when the DER is less expensive more households adopt this technology, which increases the value of a flexible demand side. Households enrolled in more complex rates alter their loads to consume electricity when it is less expensive, which is when rooftop PV systems produce it. This in turn increases the utilization of the overall fleet of (supply and demand side) generating technologies, allowing households to get more energy per unit of capacity installed. Table 5.3 shows an increase in adoption and capacity utilization when the cost of the DER is low with respect to the case when it is high.

This table also highlights the potential complementarity between DERs and time-varying rates. In this application, when the cost of the DER is low, not only dynamic rates become more valuable. The adoption of rooftop solar PV increases as rates become more dynamic. This indicates an increase in the relative value of the DER with respect to other generation alternatives. That is, time-varying rates may improve the economics of DERs.

Distributional Analysis

We explore now the results from a distributional perspective. As Figure 5.5 shows, a group of households cannot adopt the DER. This element of our setting seeks to reflect that in a population there will be some households which cannot afford adopting a distributed energy resource, such as a rooftop PV system. Now we ask the question of how the switching from a simple flat rate to a more complex structure benefits customers with different financial means—in our setting those who can and cannot adopt DERs.

Table 5.4 shows the surplus increase as a result of the switch in rate structures, for the case where

Table 5.3: Rooftop Solar PV Adoption and Generation Fleet Utilization

Network Status ¹	Cost DER	Cost newable	Re-	Tariff				
				RTP	TOU	TOU + DC	FR + DC	FR
population with rooftop solar PV [%]								
const.	low	low		24	23	23	19	16
		high		24	23	22	20	16
	high	low		7	11	15	16	7
		high		7	10	12	15	7
unconst.	low	low		24	23	23	22	21
		high		24	23	22	22	21
	high	low		0	0	0	0	0
		high		0	0	0	0	0
capacity utilization [%]								
const.	low	low		59	56	56	49	44
		high		59	56	54	51	45
	high	low		55	54	54	47	41
		high		56	52	52	47	41
unconst.	low	low		59	56	55	47	46
		high		59	56	55	45	46
	high	low		49	48	48	43	41
		high		49	48	48	43	41

¹ const. = constrained, unconst. = unconstrained.

Table 5.4: Net Surplus Increase per Household with Respect to FR [%]

Network Status ²	Cost Renewable	Household Type		Tariff			
		Bus	Can Adopt DER	RTP	TOU	TOU + DC	FR + DC
const.	low	1	yes	7.28	7.26	7.24	2.48
			no	7.29	7.26	7.25	2.44
		2	yes	10.98	10.72	10.46	4.53
			no	6.60	6.55	6.80	3.48
	high	1	yes	6.60	6.55	6.80	3.48
			no	6.60	6.55	6.51	3.48
		2	yes	10.46	10.06	9.81	5.92
			no	6.94	6.75	6.50	5.28
unconst.	low	1	yes	0.28	0.20	0.21	(0.00)
			no	0.28	0.20	0.21	(0.00)
		2	yes	1.47	1.50	1.32	(0.02)
			no	(0.06)	(0.72)	(0.74)	0.20
	high	1	yes	0.22	0.22	0.24	(0.05)
			no	0.22	0.21	0.21	(0.05)
		2	yes	1.59	1.58	1.28	(0.19)
			no	(0.14)	(1.15)	(0.74)	0.19

¹ A 10 % increase is between \$100 and \$250 per year.

² const. = constrained, unconst. = unconstrained.

the cost of the DER is low. When the switch improves welfare the most, which is when the network is constrained, the disparities are more significant. In one of the buses, for those household that can afford the DER, the increase in surplus is approximately 1.5 times that of the households that cannot afford a PV system. It is interesting to notice that this is not always the case, as there are no differences in the surplus increase between the households located in bus 1. When weather patterns and network conditions are favorable, wealthier households can take more advantage of sophisticated rates by adopting distributed energy resources. Additionally, the model indicates that the flexibility of the structures across time and states of nature makes this phenomenon more pronounced. In fact, the switch to RTP translates into surplus increases for households that can adopt the DER being 1.6 times that of households with less financial means. On the other hand, when the switch is to the FR + DC structure, the surplus increase of wealthier households is only 1.2 times that of those that cannot afford the DERs.

In Table 5.4 we can also see that in the constrained case the switch to a more complex rate is always Pareto improving. All type of households are better off after the switch. This does not happen when the network is unconstrained. In most cases all households increase or maintain their net surpluses. However, when the cost of the renewable generating technology is high and the switch is to a time of use rate, households that cannot adopt DERs and are located at Bus 2 are worse off after switching. While the absolute harm is small, approximately \$31 per year, since it happens to household with comparatively less financial resources, it potentially has more negative consequences than if it were borne by those that can afford adopting DERs. In addition, we see that in the unconstrained case time-varying rates continue to accentuate disparities. While the switch to a FR + DC rate practically has no effect on consumer surplus, it does have an effect when it is to a time-varying rate. As in the constrained case, households that can afford rooftop PV systems take advantage of this technology to increase their surpluses under the time-varying rate.

5.6 Conclusions

This study develops a technique to compare rate structures. It allows researchers to transparently model a wide range of tariffs, distributed energy resources and supply side configurations. The specific values of the tariffs, as well as the demand and supply side consumption, production and investment decisions result from solving the problem of a Ramsey-Boiteux planner. This is the problem of a regulator that anticipates the impacts of its choice of rates on demand and supply side short- and long-run decisions.

We cast this Bilevel Model as a Mathematical Program with Equilibrium Constraints, and solve it using a variant of the alternating direction method of multipliers. This variant handles the complementary constraints in a distributed manner. It solves the MPEC iteratively, at each iteration solving independently several small MPECs and one nonlinear program. Using the technique of [58] we cast the small MPECs as MIPs and develop heuristics to find reasonable starting points; we handle the nonlinear program via a conic reformulation. A computational exercise demonstrates that our solution approach has several desirable properties for practical applications. First, it vastly outperforms Knitro, a popular commercial solver for MPECs. Second, the method shows good convergence behavior, reaching solutions of reasonable quality after a few tens of iterations. Third, the algorithm scales well with size as part of the resolution of the problem can be distributed

to various computing nodes and the solution time of the centralized step increases at a rate lower than linear.

With a numerical analysis, we demonstrate the importance of the modeling flexibility of the present technique to compare rate structures. In contrast to previous approaches, our method allows capturing complex supply side configurations. In particular, researchers can model a network. Our exercise shows that abstracting away this element can have meaningful impacts on the analysis of rate structures. In the example, the presence of network constraints translates into significantly higher welfare gains when switching from a flat rate to more sophisticated structures. Omitting this element underestimates the benefits by about 3000 times. In addition, the analysis highlights the value of being able to model DERs. It shows how the economics of rooftop PV systems impact the rate comparison, and that time-varying rates and inexpensive DERs could complement each other; the former can improve the relative value of the latter while DERs being inexpensive increases the gains of switching from a time invariant rate. Finally, the distributional analysis portrays how our method permits regulators and policy makers to study impacts of a rate update on a heterogeneous population. While a switch in rates could have a positive impact on the aggregate of households, it could benefit some more than others. It could even harm some customers, which can be particularly problematic if those harmed were low income households. A technique such as the one introduced in this paper permits to anticipate these impacts, letting regulators and policy makers to decide among rates structures with considerably more information than what would be available with alternative techniques.

6. Utility Pricing in the Prosumer Era: An Analysis of Residential Electricity Pricing in California

6.1 Introduction

The increasing penetration of *distributed energy resources* (DER) including advanced metering infrastructure (AMI), energy management systems (HEMS), solar photovoltaic and battery storage systems are enabling residential consumers, or “prosumers”, to interact bidirectionally with electricity systems. Customers can not only purchase electricity from the grid but also provide the system with energy and reliability services [130]. In regulated retail sectors, residential rate structures greatly influence this interaction. Rates can impact adoption decisions, by changing the economic value of the technology, and can influence the usage of this resource. Rates may also increase distributional disparities, creating cross subsidies between those who can and cannot adopt the distributed technology [49]. Despite this crucial role, no consensus exists with respect to the ideal rate structure. Factors explaining this reality include limitations of the theory and the focus of the empirical work. While the theory asserts that real-time pricing (RTP) is optimal from an economic efficiency perspective [73], regulators must balance efficiency with other public goals, such as rate simplicity, equity or meeting environmental directives [134]. Absent a comprehensive quantification of the impacts associated with implementing RTP, it is difficult for regulators to judge the value of this alternative, especially when it may compromise other regulatory goals. The empirical literature has tried to quantify impacts, however, the scope has been somewhat limited. Researchers have focused on estimating price responsiveness to quantify changes in efficiency in the short-run.¹ Other relevant metrics such as long-run welfare effects, equity or environmental implications have been explored either in isolation or with stylized analyses, but never with an applied approach, within a unified framework.²

This work contributes to rate regulation policy with an applied study in the context of California’s residential electricity sector. Our focus is the comparison of the long-run welfare effects as well as the equity and environmental implications of a set rate structures. The analysis considers the hypothetical scenario in which HEMS is widely adopted. Under this circumstance the household responses to price signals are likely to be fully rational, driven by an algorithm optimizing the consumption of the appliances [7].

A second contribution is the development of a framework that permits comparing rate structures along each of the metrics we consider in our analysis. We develop a model of optimal pricing that accommodates a wide variety of tariffs. Our framework builds upon *peak-load pricing*³, and borrows some elements from the literature of *generating capacity expansion*.⁴ We embed a detailed model of household behavior in this setting, expanding the basic model of peak-load pricing to include heterogeneous households and the adoption of DERs. Our approach allows us to capture a wide variety of temporal and spatial demand substitution patterns, without needing to use a large number of estimates.

¹Examples include the work of [4], [29], [53], [54] and [66].

²See, for instance, [13], [11], [38] and [73].

³For a comprehensive survey of the literature we refer the reader to [38].

⁴[127] provide a good example of these models.

Our analysis is particularly relevant to the current regulatory situation in California, in which the default increasing block pricing program is gradually being retired and time-of-use becomes the default rate. We compare this prospective structure with two variants of this program, a TOU combined with demand charges (TOU&DC) and a TOU combined with a critical peak pricing program (TOU&CPP). We also include in the comparison the cases in which households enroll in a flat rate (FR) and in a real-time pricing program.

Considering all households under the FR rate as a reference case, our analysis shows that implementing any other pricing alternative produces gains for the average household that are mild at best. When implementing time-of-use pricing, or any of its variants, the average gain is not greater than 1.2 dollars per month. Even though real-time pricing performs better than the other rate structures, the improvement over the FR tariff seems mild. The average household increases its net surplus by 2 dollars per month under this rate scenario.

In addition, our analysis shows that the net surplus gain varies considerably across households. Factors such as the presence of an air conditioning system or the temperature outside the dwelling are major drivers of this variation. For all rates, households with air conditioners experience higher average gains than household without these appliances. However, the gains in net surplus vary more across the former group of households. The exterior temperature profile is a key factor. Its relationship with the net surplus gains, however, is not simple, and depends on the specific rate. For instance, under the RTP case customers experience greater gains in areas with higher average temperature. But this statistic has no correlation with gains under the TOU&DC program.

These two results combined suggest that defaulting all residential customers into a time-of-use rate structure, which is the current path California is following for the residential sector, may not be the best strategy. Targeting different rates to households with different appliance stocks and in different locations will likely be a superior policy.

6.2 California Electricity Sector and the Emergence of Prosumers

6.2.1 An Overview of the Sector

The California electricity sector serves approximately 30 million people across the state. With 59 GW of power plant capacity, the sector delivers near 309 TWh of electricity annually. Its market size is close to \$8 billion per year and its transmission system, spanning 25,627 circuit-miles, is part of the Western Interconnection. In terms of market regulation and oversight, there are three institutions involved, each with different roles. The first, the Federal Energy Regulatory Commission (FERC), has jurisdiction over the interstate transmission of electricity. Its responsibilities include the oversight of important merger and acquisitions, reviewing applications for transmission projects, as well as licensing and inspecting private, municipal and state hydroelectric projects. The commission also sets mandatory reliability standards and monitors energy markets across the US. The other two regulatory agencies have jurisdiction in the state of California. One is the California Energy Commission which is the primary energy policy and planning agency of the state. The other is the California Public Utilities Commission (CPUC). Its main role is regulating the three investor-owned electric utilities of California, including Pacific Gas and Electric Company (PG&E), Southern California Edison (SCE) and San Diego Gas and Electric Company (SDG&E),

which collectively serve two thirds of the electricity demand throughout California. Among other functions, the CPUC sets and approves the retail rates, is responsible for ensuring that utilities meet state environmental policies and ensures electricity safety at the distribution level.

In terms of system and market operations, the California's Independent System Operator (CAISO) is responsible for maintaining a reliable transmission of power as well as the comprehensive long-term planning of grid infrastructure. This entity also coordinates forward and spot markets for energy and ancillary services. In addition, CAISO complies with the reliability standards set by the North American Electric Reliability Corporation (NERC) and the Western Electricity Coordinating Council (WECC). While the former is a non-profit organization developing and enforcing reliability standards for the continental United States, Canada and Baja California, Mexico, the latter is a regional entity promoting bulk electric system reliability in the Western Interconnection.

6.2.2 Residential Rates in California

The distinctive characteristic of residential electricity rates in California is its increasing block structure. They have had this form since 1976, when the Miller-Warren Energy Lifeline Act was enacted. This legislation sought to provide California's residential customers with a minimum necessary quantity of gas and electricity at a fair price, and also to encourage conservation. The legislation set a precedent, providing a conceptual justification for implementing increasing block rates. Since 1976 rates did not change meaningfully until California's electricity crisis.

Beginning in the summer of 2000, tight supply margins, weak federal oversight, lack of an elastic demand and flaws in the market design yielded a period of highly volatile electricity prices, known as California's Electricity Crisis. As a result of the crisis the sector underwent a period of drastic reforms. At the retail level, a first response to the high wholesale prices was to lift the retail price cap. This triggered notorious increases in electricity bills, which were then mitigated by freezing the charges of the lower two tiers. The result of this legislation was the replacement of a two tier system by a five tier structure, with prices of Tiers 3 to 5 considerably higher than those of the remaining lower tiers. From 2000 to 2009 differences among tiers increased. However, the enactment of SB695 in 2009 began to allow limited annual increases for Tiers 1 and 2.

At the time of this writing, decision *D.15 – 07 – 001* is the main piece of regulation laying the path for the future of residential electricity rates in California. Key elements of this regulation include the promotion of the consolidation of the tiers and the development of rates that reflect better cost causation. In particular, the decision approves transitioning all residential customers to a default time-of-use tariff by 2019.

6.2.3 The Emergence of Prosumers

There are two main forces pushing the emergence of prosumers in California: Environmental policy directives and distributed technologies reaching maturity. The relevant environmental policy is the renewable portfolio standard which established, among others, targets for distributed generation. The policy has triggered the development of an array of incentives for generation at the customer's premises which has caused the massive deployment of distributed energy resources,

such as solar photovoltaic panels. As for technological evolution, California has taken major steps towards modernizing its distribution grid, having the largest installation of AMI in the US. This technology constitutes a vital element for implementing time-varying-rates. By enabling two way communication between customer and utility on time intervals of an hour or less, AMI allows utilities measuring consumption on an hourly basis as well as sending price signals on a consistent time scale.

6.3 A Modeling Framework to Compare Rate Structures

6.3.1 Utility Pricing: An Overview of the Theory

An important function of regulators is determining the rates that a regulated utility can charge for the provision of its services. This process, also known as *rate regulation*, includes the determination of the rate *rate level* and the *design of rate structures* [115, pp. 176 - 180]. Establishing the rate level entails specifying the total compensation that the utility, or load serving entity (LSE), receives for its services. The design of rate structures, which is the focus of this work, defines how the LSE collects its compensation. Designing rate structures is far from trivial. So it is not surprising that a myriad of methodologies have been suggested and adopted in different jurisdictions. [20] divide the approaches according to how they assign *common* or *non-attributable* costs across different services and consumer classes. There are two broad categories: the *cost-based pricing* and pricing based on the concept of *marginal cost*. The first group of approaches allocate costs based on criteria other than efficiency. One example is the *fully distributed costs method*, in which common costs are assigned according to the relative shares of magnitudes that can be attributed to a service or group of customers, such as peak-demand, output or revenue.⁵ On the other hand, in pricing based on the concept of marginal cost efficiency has a prominent position. How common costs are attributed to the different groups of customers is a byproduct of a welfare maximization process. Our framework falls into the second category of approaches. The model that this work introduces produces a set of rates and allocations of costs that emerge from the welfare maximization of the system under study. This is the approach to rate design that the literature of peak-load pricing studies.

Peak-load pricing develops a normative theory of efficient or welfare maximizing pricing for industries with limited storage capability and time-varying demand. The modern version of the theory originates with the contributions of [9] and [135], and intended to provide guidelines in the context of price regulation of natural monopolies, such as vertically integrated electric utilities [38]. The basic model considers the problem of a social planner choosing prices that maximize welfare, i.e., the surplus of customers and the public utility's profits. Prices coordinate production and consumption decisions over a time horizon. The monopolist invests in production capacity at the beginning of the horizon and prices are such that the utilization and the level of the installed capacity are optimal [46]. Studies including [25], [31], [39] and [109] further refine the model to include a stochastic demand, supply-side uncertainties and multiple technologies.

⁵Other approaches seek to minimize cross subsidies across services and consumer classes and others build a set of axioms and derive rate structures consistent with them. For more details on the subject of cost-based pricing, see [20, pp. 44 - 60].

[15] and [73], in examining the merits of retail competition in the electricity industry, show that the theory also applies to restructured electricity sectors. In these models, a competitive wholesale market replaces the production side of the vertically integrated utility. At the retail level, both papers distinguish the cases of a regulated distribution company and competitive retailers. While [15] consider a setting with linear or uniform prices, [73] contemplate the case of a two-part, non-linear price. Further, [15] explore the long-run effects of different pricing policies, analyzing the equilibria that emerge at the wholesale and retail levels. More recently, [159] investigates a setting in which there is imperfect competition at the wholesale level, and [32] updates earlier work exploring the interaction of different pricing policies and renewable technologies.

Our model departs from previous work first by generalizing the type of rate structures present in peak-load pricing. In addition, we introduce a mechanism linking pricing and technology adoption decisions. Finally, our model accommodates household heterogeneity beyond a scale factor.

6.3.2 The Regulator's Problem

The regulator's problem combines elements of the peak-load pricing and capacity expansion literature. The key element from capacity expansion not present in peak-load pricing is a transmission network. For simplicity, we do not detail this element of the model in this section. As in peak-load pricing, our model falls into the broad category of two-stage stochastic optimization models. Agents in these models make long-run decisions at the beginning of the horizon before uncertainty is realized and define state contingent strategies for the short-run stage. These are static models that can describe systems in steady state. Our framework, therefore, is not suitable for studying system dynamics. In terms of the institutional setting, at the retail level we consider a distribution utility as the load serving entity. In general, however, one can consider settings within two polar cases. While the utility could be fully integrated with the supply side in one case, in the opposite, it could be just a distribution company. Under the assumption of perfect competition at the wholesale level, both cases are equivalent [74], however.

Let ω index a finite and countable set Ω of states of nature, π the corresponding probability vector, $E[\cdot]$ the expectation operator, and a time horizon of $t \in T$ time steps. Given a random vector of consumption d , the household pays $l + \eta(d, p)$ to the utility, where l is a fixed charge, $p \in \mathcal{P}$ a vector of rate parameters and $\eta(\cdot)$ a fee contingent on consumption and the rate parameters. We call the triple $(l, \eta(\cdot), p)$ a rate structure. Our setting is similar to the one in [73] insofar we focus on two-part structures with a state and time contingent demands and prices. However, we generalize this model to accommodate more complex rate structures. In our setting the vectors d do not only have one component for every time and state of nature but also may include other relevant metrics associated to the demand profile, such as peak or total consumption across the time horizon. Similarly, price parameters may include charges for peak or total demand. Specifically, we focus on the case in which $\eta(\cdot)$ is bilinear on the demand vector and price parameters. The following assumption formalizes this specification.

Assumption 3. *The demand contingent charge $\eta(d, p) = d^\top M p + \text{Ind} \left\{ \tilde{b} - \tilde{A}d \right\}$, where $\text{Ind} \{x\}$ is 0 if $x \geq 0$, and ∞ otherwise.*

As subsection 6.3.5 shows, this specification is fairly general, allowing researchers to model a wide

range of the rate structures used in practice.

As in peak-load pricing, we consider homogeneous households, with a mapping $D_\omega(p) : \mathcal{P} \rightarrow \mathbb{R}^{|T|}$ and a real valued function $U_\omega(d_\omega)$ representing their demand and gross surplus metric, respectively. Given a set of wholesale prices $\{\lambda_\omega\}$, the planner problem optimizes the household net surplus, $E[U_\omega(D_\omega(p)) - \eta(D_\omega(p), p)] - l$, and guarantees that the utility meets its revenue requirement, $E[\eta(D_\omega(p), p) - D_\omega(p)^\top \lambda_\omega] + l - \Pi$, with Π an exogenous fixed cost. As [73] shows, this amounts to find (l^*, p^*) such that

$$(d^*, p^*) \in \arg \max_{(d, p)} \{E[U_\omega(d_\omega) - d_\omega^\top \lambda_\omega] : d_\omega = D_\omega(p) \ \forall \omega \in \Omega\}, \quad (6.1)$$

$$l^* = E[\lambda_\omega^\top d_\omega^*] - \eta(d^*, p^*) + \Pi. \quad (6.2)$$

6.3.3 A Competitive Wholesale Electricity Market

The wholesale market representation in this model is a variant of the supply-side model studied in the peak-load pricing and capacity expansion literature. More specifically, we follow closely the representation in [32]. In this model, infinitesimal competitive firms interact in a spot market for electricity. Each decides on their long-run installed capacity and short-run generation profiles. We denote the total installed capacity of technology $k \in K$ as x_k and its cost of carrying capacity as \tilde{r}_k . The aggregated production profile of this technology in state of nature ω is $y_{\omega k} \in \mathbb{R}_+^T$, and variable costs per unit of power production is $c_{\omega k} \in \mathbb{R}_+^T$. We capture variability in a technology's availability – e.g. due to outages – with an availability factor per technology contingent on the states of nature, $\rho_{\omega k} \in R^T$. In a perfectly competitive market firms are price takers, thus, production and capacity for technology k are the solution of the problem

$$\max_{(y_{\omega k}, x_k)} \{E[(\lambda_\omega - c_{\omega k})^\top y_{\omega k}] - x_k \tilde{r}_k : 0 \leq y_{\omega k} \leq x_k \rho_{\omega k}\}. \quad (6.3)$$

The market equilibrium is a tuple $(d^*, p^*, y^*, x^*, \lambda^*)$ such that (d^*, p^*) solves the regulator's problem at λ^* , (y^*, x^*) solves the problem of the producer at that price, and supply equals demand. It is easy to verify that the market equilibrium is the solution of

$$\max_{(d, p, x, y)} E \left[U_\omega(d_\omega) - \sum_{k \in K} y_{k\omega}^\top c_{k\omega} \right] - x^\top \tilde{r} \quad (6.4)$$

subject to

$$d_\omega = \sum_{k \in K} y_{k\omega} : \lambda_\omega, \quad (6.5)$$

$$0 \leq y_{k\omega} \leq x_k \rho_{\omega k}, \quad (6.6)$$

$$p \in \mathcal{P}, \quad (6.7)$$

$$d_\omega = D_\omega(p). \quad (6.8)$$

6.3.4 The Household Behavior

Except for our specification of the demand contingent fee, $\eta(\cdot)$, (6.4)–(6.7) is the classic peak-load pricing problem. Researchers can use the model to analyze theoretically and numerically implications of different constraint sets for the vector of retail prices. A key assumption that facilitates the study of these models is a demand system with analytic expression. Our framework drops this assumption because our specification of $\eta(\cdot)$ implies, in general, demands with no analytic definition. Consistently, our model updates the peak-load pricing problem replacing (6.8) with the following condition,

$$d \in \arg \max_d \left\{ E \left[U_\omega(d_\omega) - d_\omega^\top M_\omega p_\omega \right] : \bar{b}_\omega - \bar{A}_\omega d_\omega \geq 0, \forall \omega \in \Omega \right\}, \quad (6.9)$$

where (\bar{b}, \bar{A}) contain the parameters of the rate structure, (\tilde{b}, \tilde{A}) , and possibly others. Henceforth we refer to (6.9) as the *household problem*, and to (6.4) – (6.7), (6.9) as the *pricing problem*.

6.3.5 Illustrative Examples

Our specification of the household demand allows to model the influence on demand of several rate structures and, also, represent demand-side technologies of interest. Here we show how to implement the models that we use in our analysis. Some notation will prove useful. The matrix I_m corresponds to the identity of m by m . The vectors e_m and z_m are, correspondingly, vectors of ones and zeros of m dimension.

Modeling rate structures. Our analysis compares *time-varying pricing* (TVP) and a TVP combined with a *demand charge* (DC). A time varying pricing is the simplest type of rate to model. Set $M_\omega = I_{|T|}$, let the vectors $p_\omega \in \mathbb{R}_+^T$ and $d_\omega \in \mathbb{R}^T$, and define \mathcal{P} as follows,

$$\mathcal{P} := \begin{cases} \left\{ p \in \mathbb{R}_+^{|T| \times |\Omega|} : p_{\omega t} = p_{\omega' t'} \forall (\omega, t), (\omega', t') \right\} & \text{for FR,} \\ \left\{ p \in \mathbb{R}_+^{|T| \times |\Omega|} : p_{\omega t} = p_{\omega' t'} \forall (\omega, t), (\omega', t') \in TW(\omega, t) \right\} & \text{for TOU,} \\ \left\{ p \in \mathbb{R}_+^{|T| \times |\Omega|} \right\} & \text{for RTP,} \end{cases} \quad (6.10)$$

where $TW(\omega, t)$ is the set of time windows (ω', t') in the same time window as (ω, t) .

Adding a demand charge to any of these structures requires redefining $d := [\bar{d}, \hat{d}]$ and $p := [\bar{p}, \hat{p}]$, where $\bar{d}, \bar{p} \in \mathbb{R}_+^{|T| \times |\Omega|}$, and $\hat{d}_\omega, \hat{p}_\omega$ correspond to the maximum consumption and demand charges under ω , respectively. The matrix M_ω is now equal to $I_{|T|+1}$, and the analyst may add additional conditions to the set \mathcal{P} to model demand charges constant across some scenarios. A final element of this structure is the constraint linking the hourly consumption profile \bar{d} and the maximum consumption \hat{d} , which we model via the following definitions

$$\tilde{b} := z_{|\Omega| \cdot |T|}, \quad \tilde{A} := [I_{|\Omega| \times |T|} \quad -I_{|\Omega|} \otimes e_{|T|}]. \quad (6.11)$$

Household as composite of devices. Following the approach of [122], we consider that households are composite of devices and assume their utility functions are additively separable. For

reasons we explain later, we consider in our analysis two devices, a central air conditioning unit and a rooftop solar panel, and the household baseline. Central air conditioning falls in the more general category of *thermostatically controlled loads* (TCL's), whose behavior follows the laws of thermodynamics. [99] presents a model describing the behavior of these appliances, which links the household's inside temperature, θ , with the outdoor temperature $\tilde{\theta}$, the thermal characteristics of the dwelling, ξ , and the electricity consumption of this appliance, d . It is possible to show that the inside temperature profile has the following form

$$\theta(d; \xi, \tilde{\theta}) = \Theta_1(\xi)d + \theta_2(\xi, \tilde{\theta}), \quad (6.12)$$

where Θ_1 and θ_2 are a matrix and a vector, functions of the thermal parameters and temperature outside the dwelling. Here we close the model of the TCL behavior introducing a mechanism capturing household's preferences for thermal comfort. The simplest approach involves a penalty for deviating from an ideal inside temperature, $\hat{\theta}$. Equation (6.13) shows a utility function consistent with this approach, which we use in our analysis.

$$U(d) = -\beta \|\theta(d; \xi, \tilde{\theta}) - \hat{\theta}\|^2 \quad (6.13)$$

For modeling the household baseline, we assume a linear demand system and compute the associated utility function using a standard procedure. For the rooftop solar panel we add to the household problem a constraint limiting its hourly production given the hourly availability of the solar resource. A final element of the household model links the demands of each device with the net demand of the customer,

$$\bar{d} = d_{baseline} + d_{ac} - d_{solar}. \quad (6.14)$$

6.3.6 Household Heterogeneity and DER Adoption

The pricing problem contemplates one representative customer and there is no mechanism modeling customer adoption. In our analysis, however, we consider heterogeneous customers and analyze impacts of pricing on adoption. We incorporate these two elements using the framework that [27] develop. The paper distinguishes different customer types $i \in I$,⁶ each of which decides a set of technologies to adopt $j \in J$.⁷ Calling the combination $h := (i, j)$ a *segment* and defining α_h and r_h , respectively, as the number of households and cost associated to a segment,⁸ the paper shows how modifying the pricing problem permits modeling adoption decisions. Specifically, equation (6.4) becomes

$$\max_{(\alpha, d, p, x, y)} E \left[\sum_h \alpha_h [U_{h\omega}(d_{h\omega}) - r_h] - \sum_{k \in K} y_{k\omega}^\top c_{k\omega} \right] - x^\top \tilde{r} \quad (6.15)$$

⁶In order to fix ideas consider the following two examples: $I = \{\text{with central AC, without central AC}\}$ or $I = \{\text{live in hot weather, live in cold weather}\}$.

⁷An instance of this set is $J = \{\{\text{solar PV, battery storage}\}, \{\text{solar PV}\}, \{\text{battery storage}\}\}$.

⁸The cost associated to a segment (i, j) is the annualized capital cost of the set of technologies j .

and (6.5) updates to

$$\sum_h \alpha_h d_{hw} = \sum_{k \in K} y_{wk} : \lambda_w. \quad (6.16)$$

A final element to include for modeling adoption is the feasible region for α , which ensures that the number of households per segment is consistent with the number of households per customer type.

6.3.7 Comparing Rate Structures

Subsections 6.3.2 to 6.3.6 develop an analytic tool to compare rate structures. Researchers can explore the effects of different structures on welfare and other metrics by changing the specification of the consumption contingent fee, $\eta(\cdot)$, solving the pricing problem and comparing the metrics of interest. While the method is not intended to predict what would happen were the tariff under analysis in place, it provides a consistent assessment of the potential differences between them.

Solving the pricing problem is not straightforward. The problem falls into the broad category of *Bilevel Problems*, in which a leader – in our setting, the regulator – indirectly controls the actions of the follower – the household – changing one or more parameters of her problem. Bilevel problems are hard to solve in general, and state of the art solvers can only handle problems of moderate size. For large instances, researchers have to devise specialized algorithms. Given the size of the instance we explore in this study, we had to develop a specialized algorithm as well. However, the development of the algorithm is beyond the scope of this work. It constitutes a completely separate research effort, which [28] describes in detail. The basic idea is to decompose the pricing problem into one problem per household and state of nature, and one problem that coordinates the demands of the households. The algorithm iterates solving all problems at each repetition and stops when consecutive solutions do not change. The key aspect of the algorithm is its distributed nature which, by enabling its implementation in cluster computing facilities, makes our modeling framework practical.

6.4 Modeling California’s Electricity Sector

We construct our model of the California electricity sector supplementing the network model that [118] developed for market analysis. The model consists of a network with 240 nodes, or buses, which corresponds to a topological reduction of the transmission system encompassing the Western Interconnection.⁹ This reduced system also provides generation technologies and demands at each node, and the physical characteristics of the network, including transmission constraints. The generating power plants correspond to aggregations per type of fuel. Non-dispatchable generating technologies¹⁰ such as solar or wind generation, and reservoirs such a geothermal or hydro power plants come with a year of hourly energy production. Fossil fuel technologies, on the other hand, only include physical and short-run economic parameters, such as heat rates and fuel costs. As

⁹We refer the reader to the Western Electric Coordinating Council website for detailed information on the interconnection.

¹⁰Plants with outputs that are determined to great extent by exogenous factors such as weather conditions.

for the demands, the network model includes a year of hourly energy consumption at nodes of the network corresponding to demand centers.

Before describing how we complete this data set, we make two clarifications. At first sight, the network model has more information than this analysis requires, because the Western Interconnection includes more states than just California. Using the full interconnection model, however, allows us to produce realistic import and export flows and provides the opportunity to study the impacts of residential pricing policy outside California. Because the main computational difficulties emerge from our detailed modeling of the demand, and because we do not model in detail demand at nodes outside California, the full network model did not increase significantly the computational complexity of our analysis.

A second clarification relates to our treatment of the Western Interconnection as an integrated market. That is, in our model the physics of the transmission lines is the unique factor limiting the flow of electricity through the interconnection. In practice, the administration of this system is divided among 38 *balancing authorities*, each controlling one portion of the network, and whose central role is to guarantee the reliable operation of their respective sub regions. This adds additional limitations to the flows of electricity which our model does not capture. At the time of this writing, however, California is leading the efforts to assess the impacts of a multistate regional market for the Western Interconnection.¹¹ Thus, an integrated market is plausible for the future of the interconnection.

6.4.1 Generating Technologies

Because we are interested in studying long-run impacts, we replace the cost functions in the network model with the functions we described in subsection 6.3.3. A fixed and variable cost implement these functions. The fixed cost includes the annuity associated with developing and installing the generating technology and the fixed O&M costs. The variable cost, on the other hand, encompasses fuel and variable O&M costs.

In addition to these economic parameters, technologies have associated emissions and availability factors. While the former captures the fact that different fuels have different GHG emissions, the latter reflects the fact that power plants experience unplanned outages. Emissions factors as well as the economic parameters of the generating technologies come from [48], and table 6.1 summarizes the specific values we use in this study. As for availability factors, we use the magnitudes that NERC makes publicly available through its Generating Availability Data System (GADS).

We include in the model the existing plants, by technology, for each node. As mentioned, investment decisions to expand this fleet are endogenous to the model. We treat hydro power generation as exogenous because a correct treatment of this technology, which involves the stochastic dynamic optimization of reservoirs, is beyond the scope of our model. Even though solar and wind are non-dispatchable generators, we use the time series in the data set only to compute hourly availability factors.¹² The actual hourly production for these technologies is ultimately determined by their

¹¹ See [19] for further detail.

¹² Hourly availability factors are the ratio between the hourly production and the nameplate capacity of the technology in the data set.

Table 6.1: Generating Technologies Costs and GHG Emissions

technology	variable cost	fixed cost	emissions factor
	[2015 \$/MWh]	[2015 k\$/MW-year]	[tCO ₂ eq/MWh]
nuclear	11	448	-
biomass	35	424	1.93
coal	32	361	0.65
geothermal	-	341	-
solar	-	192	-
wind	-	184	-
gas adv CC	39	84	0.33
gas conv CC	42	77	0.35
gas adv CT	63	52	0.52

installed capacity, an outcome of our model.

6.4.2 Developing a Model of California’s Residential Demand

We construct a model of the residential sector of California calibrating our model of household behavior at each node of the network. Households can consume and produce electricity on an hourly basis, and impact the system via their net demands. In terms of consumption, we consider two major categories of household end uses: cooling and non-cooling. We take this approach for two reasons. In California central air conditioning is a major source of electricity consumption, and approximately one out of every two households has this type of appliance [108]. In addition, studies at the appliance level report air conditioning to be a major source of demand responsiveness [99, 122].

The second reason relates to the hourly demand data available for this study. In our framework the baseline of a household is the fraction of its demand not modeled as any particular device. In order to calibrate this function one needs an intercept, i.e., a time series of electricity consumption and the prices in effect when this happened. For the demand part of the intercept we use the load shapes developed by Itron Inc., described in [151]. This data set disaggregates residential consumption into space conditioning and other loads.

Utility functions summarize household preferences for each end use. We calibrate them using the appliance level elasticities and marginal effects estimates that [122] report. We model the baseline consumption as a linear demand system and use an elasticity of -0.08 , corresponding to the estimate for households with no space conditioning. For the price intercept, we follow the procedure that [15] describe, assuming the rate structure of the intercept to be a flat rate.

The model for cooling corresponds to the TCL model developed in subsection 6.3.5. This has three groups of parameters which can be categorized as technical, behavioral and weather related. Technical parameters include the thermal resistance and capacitance of the household, and the efficiency of the air conditioner. [99] provide ranges for these parameters in California, and we use the midpoint of those ranges in their study. The behavioral parameters are the ideal interior

temperature, which we set to 22°C (or 72°F), and the discomfort penalty, β . Using our model of the TCL, we link this latter parameter with the estimate of the marginal effect for central air conditioning of [122]. The expression linking the two magnitudes is

$$\beta = -\frac{e_{|T|}^\top \Theta_1(\xi)^\top \Theta_1(\xi) e_{|T|}}{\frac{d}{dp} E[e_{|T|}^\top d\omega]}. \quad (6.17)$$

The weather related parameter corresponds to the outside temperature. The National Renewable Energy Laboratory (NREL) develops the Typical Meteorological Year (TMY) data set for modeling energy conversion systems.¹³ This data set contains 12 months of hourly data at selected locations across the US. The data for each month typifies conditions for the location over a longer period of time, such as 30 years. There are 73 locations corresponding to California. Based on distance we assign each of these locations to one of the buses with demand within the California portion of the network model. The TMY data set also comes with hourly values for solar radiation. We use them for our model of rooftop solar panels.

Another input for the analysis is the count of households types per bus. We distinguish four types of households in our analysis, corresponding to the combinations of tenure status and the presence of central air conditioning. While section 6.5 discusses this categorization further, here we focus on the calibration of the household counts. The sources of data for this task are the 2010 census and the Residential Appliance Saturation Survey 2010 (RASS), described by [108]. The General Housing Characteristic data set of the 2010 census contains counts of occupied households by tenure. Based on distance we assign census counts at the tract level to each bus in order to calibrate the total number of renter and owner occupied households. Similarly, we use the distance between the zip code centroid and the bus to assign the RASS responses to each bus. The survey, besides recording appliance ownership per survey participant, includes their tenure status. We assume that the fraction of household under each of the household types is that of the RASS survey and we multiply this fraction by the census counts per bus to estimate the final number of households under each of the categories that we analyze.

A final piece of the demand side is that corresponding to commercial and industrial customers. The test system comes with total load profiles per bus. The commercial and industrial load at each node of the network model is the difference between the total and aggregated residential demand at each node. We assume the latter quantity to be equal to the baseline profiles multiplied by the households counts.

6.5 An Analysis of Residential Rate Structures in California

This analysis explores efficiency, as well as the distributional and environmental impacts of residential electricity rate structures. We use our model to quantify these metrics for five different tariffs that constitute plausible future residential rates in the Californian electricity sector.

Because our analysis focuses on long-run impacts, ideally one would have to compare net surplus distributions with respect to wealth levels under the different pricing regimes in our study. This

¹³For a detailed description of the data we refer the reader to [152].

approach is impractical, however, for at least two reasons. First, to the best of our knowledge information on households wealth for California is not publicly available. Thus, having this input would require an indirect calculation, which is an effort beyond the scope of this research. A second reason is that our model does not directly account for household wealth. There is no explicit mechanism linking this metric with either short- or long-run decisions. We assume, alternatively, that the level of wealth of a household translates into differences in its technology options. Wealthier customers have access to a wider variety of technologies.

Consistently, we split the population according to whether they can or cannot adopt distributed energy resources. The splitting criteria is household tenure status. That is, we assume homeowners have enough resources to purchase DERs while renters do not. Even though this criteria reflects the reality in California in the past decade [14], with better financing alternatives and new business models such as community solar¹⁴ our assumption may not be adequate for future analyses.

An additional clarification relates to the specific rate structures we use in our analysis. An important element determining the final definition of the time-of-use schedules are the time windows associated with the different charges. Commonly, these rate structures distinguish valley and peak periods and also seasons in order to set the volumetric charges. The traditional approach is to consider existing time windows as inputs. In this analysis we take a different path which avoids two difficulties involved with the traditional approach. One is that existing time windows are likely inadequate for future system conditions. In California only a small fraction of households have been enrolled in TOU's programs. As TOU becomes the default rate for residential customers and the *net load*¹⁵ shape changes due to the increasing penetration of renewable generation, the existing time windows will likely be obsolete. Furthermore, a TOU with a demand charge rate has not yet been implemented at the residential level in California. A second difficulty is the sub-optimality of setting time windows exogenously. This makes the comparison among rates inconsistent because our framework computes optimal retail prices for the RTP and FR programs.

In order to avoid these shortcomings, we consider the most flexible type of TOU possible. That is, tariffs in which the energy charge can vary hourly and across seasons but not for days occurring in the same season. Similarly, we assume a demand charge that changes across seasons. Our preliminary analysis indicates that four to five time-windows, depending on the season, can approximate with no meaningful efficiency loss the hourly windows. The results we report in this analysis, however, correspond to the hourly energy charges.

A final clarification relates to the DERs we include in our exercise. Besides considering all residential customers having AMI and HEMS, originally, homeowners were able to adopt either rooftop solar PV systems or battery storage units. Preliminary results indicated that the latter two DERs were not cost-effective alternatives. No customer under any of the rates we study, nor under current or projected economic parameters for these technologies, adopted these DERs. In the case of the solar PV systems, this result indicates that the factors driving the current adoption levels of this DER are policies specially designed to promote this technology. These include the California Solar Initiative (CSI), federal subsidies and the increasing block rate structure for residential customers. Because the CSI is not effective anymore, and the future of federal subsidies is uncertain, we do not include these policies in our analysis. On the other hand, even though increasing block struc-

¹⁴We refer the reader to [70] for further discussion.

¹⁵Net load is the net of the aggregated demand, or system load, and the non-dispatchable generating technologies.

tures are being phased out, a surcharge for high monthly consumption will remain. This makes the study of this structure relevant. Future versions of this analysis will include this rate.

In the case of battery storage systems, our preliminary results do not make a case against this technology. It simply reflects the fact that our model only accounts for the energy arbitrage value that a battery storage unit can create. In addition, however, this technology can provide ancillary services for the distribution grid and serve as a mean of transportation when being part of an electric vehicle. Because our model does not capture any of these value streams, we do not include this technology in the final analysis.

6.5.1 Aggregated Efficiency Gains

Table 6.2 shows efficiency gains of four tariffs with respect to the base case scenario: the flat rate structure. At the level of the Western Interconnection, the RTP rate achieves greater efficiency gains followed by the time-of-use combined with a critical peak pricing program. The TOU combined with a demand charge produces similar gains but the time-of-use program alone only increases the net benefit by one half of the value when combined with another program. All programs reduce the aggregated benefit - or gross surplus - of the residential sector. However, the reductions in costs more than compensate the reductions in gross surplus.

In terms of efficiency increases, the same ranking does not hold when focusing on the residential sector in California. The main difference is that the TOU&DC rate structure increases the net benefit the least. This is the result of differences in bill reductions for customers inside and outside California. The presence of a transmission network explains this outcome. In our framework, the bill of a customer is equal to the multiplication of the *locational marginal prices* (LMPs)¹⁶ by her consumption profile. The topology of the network significantly influences the magnitude of the LMPs at different nodes and, thus, the household bill at different locations. Differences in LMPs then explain differences in the distribution of bill reductions, in and outside California. In the case of the TOU&CPP rate, customers outside California capture an important fraction of the cost reductions with respect to the FR case. In all the other cases, on the other hand, the Californian residential sector captures most of the reductions in costs.

The average efficiency gains per household are mild at best, being not greater than 2 dollars per month. Importantly, in all cases, with the exception of the RTP program, the gains appear insufficient to justify the implementation of time-varying rate structures. Implementing any time-varying rate requires the deployment of AMI. Estimates of the cost of this infrastructure vary. However, one can construct a reasonable range using the documentation of pilot projects conducted under the American Recovery and Reinvestment Act of 2009 in [43, 45]. Considering the cost of AMI, the net of the average expenditure per household on advanced metering infrastructure and the operational savings, plus the cost of a standard meter, a reasonable approximation of this cost lays between 1 and 2.5 dollars per month. The lower bound at least doubles the gains of TOU and TOU&CPP, warning against the deployment of AMI if these tariffs are the pricing alternatives. Even though in California AMI is already deployed, unless the cost of AMI decreases, in the long

¹⁶In many jurisdiction, in particular in California, the wholesale electricity prices differ at different nodes of the transmission network, reflecting network congestion and transmission losses. These nodal prices are called locational marginal prices.

Table 6.2: Benefits and Costs: Changes with Respect to Flat Rate Structure

Level	Tariff	Net benefit	Benefit	Cost	Net benefit as a percentage of the cost
		[millions \$/year]			[%]
Western Interconnection	RTP	340	-100	-440	1.50
	TOU & CPP	155	-38	-193	0.68
	TOU & DC	137	-22	-159	0.61
	TOU	77	-22	-99	0.34
California's residential sector	RTP	274	-100	-374	5.18
	TOU & CPP	46	-38	-84	0.87
	TOU & DC	176	-22	-198	3.28
	TOU	82	-22	-104	1.53
[\$/year]					
Average per household in California	RTP	22	-8	-30	5.18
	TOU & CPP	4	-3	-7	0.87
	TOU & DC	14	-2	-16	3.28
	TOU	6	-2	-8	1.53

run the state might do as well with simpler rates and a simpler infrastructure.

6.5.2 Implications for Different Households

Figure 6.1 shows the distribution of households across net surplus gains with respect to the flat rate tariff. The figure has four panels, one per each type of household we distinguish in this analysis. In terms of the average gain, the ranking that we observe at the household level in Table 6.2 also holds when disaggregating per type of household. While RTP remains the most beneficial rate structure, the combination of a TOU and critical peak pricing program is the least favorable. Even some household types would be better off with a simple flat rate tariff than with the TOU&DC structure.

For all rates, the average net surplus gain is different for different households. Those with central air conditioning have a greater average surplus gain when compared to households without this appliance. This difference translates in turn into homeowners having a greater average surplus gain than renters. This happens simply because the proportion of homeowners with central AC is greater than 50%, and the opposite is true for renters. If one consider home-ownership as a proxy for wealth, wealthier customers benefit more from the rate structures we explored in this analysis.

Customers with no AC systems experience small average net surplus gains and this metric has small variance across households. The small increase in net surplus is driven by the elasticity we assume for the baseline consumption. The demands of households with no AC system are inelastic.

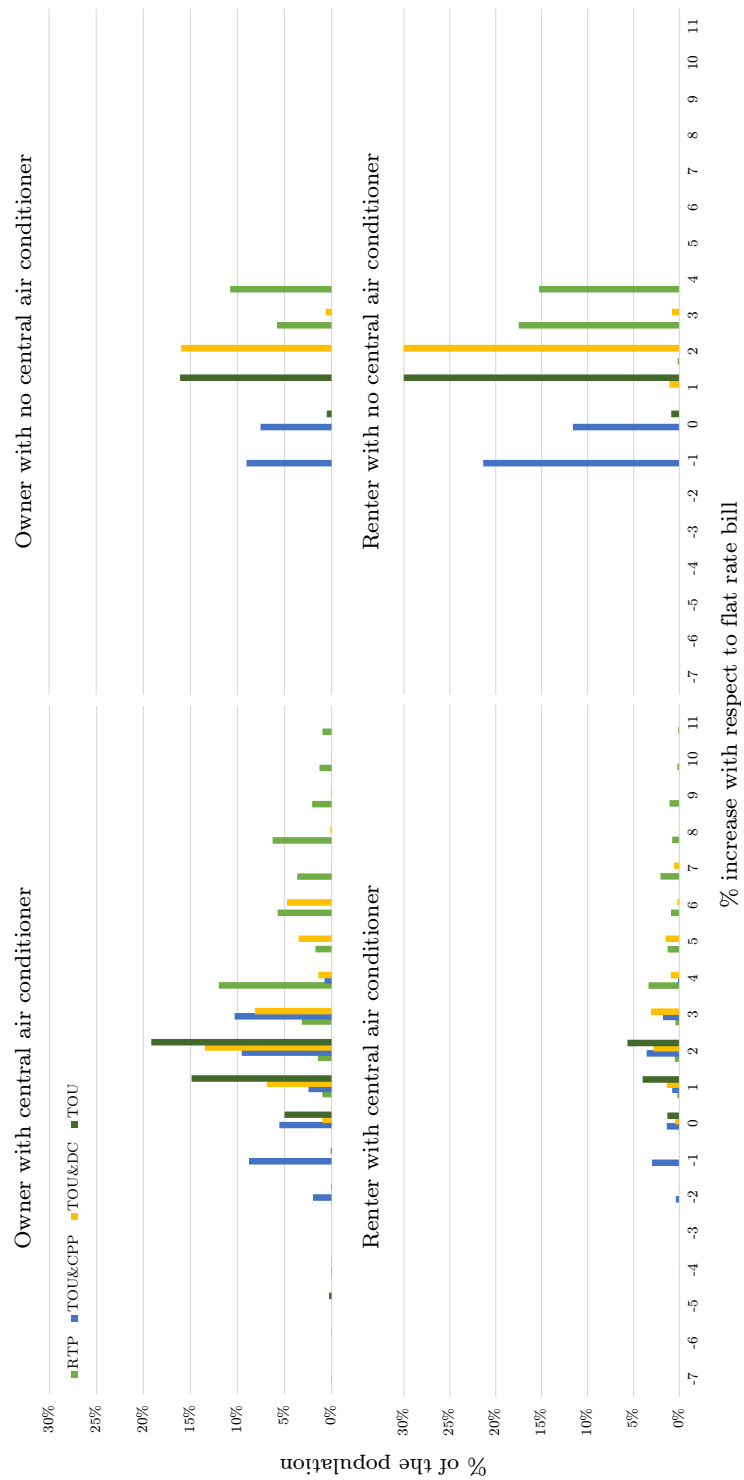


Figure 6.1: Distribution of Households Across Net Surplus Gains

This small elasticity helps to explain the small variance as well. In addition, other two elements influence the variance. One is the fact that we use one baseline profile for all households. Another is that the LMPs show small variation in the demand nodes of the California portion of the network.

The net surplus gains vary more across households with central AC systems. The variance is driven by differences in the temperature profiles at the different locations. Interestingly, the order in terms of net surplus gains induced by the different temperature profiles is different for each rate we explore. For instance, in the case of real-time pricing locations with higher within-day temperature variance and higher average temperature tend to have greater net surplus gains. One observes a similar pattern when households are enrolled in the TOU&CPP and TOU programs. However, this correlation disappears when the time-of-use rate is combined with demand charges. In this latter case, households located in places where the between-day variance of temperatures is lower tend to benefit the most.

The variance in net surplus for household with AC systems suggest targeting as a strategy for implementing time-varying rates. In particular, the TOU&DC and the RTP program appear to be the most attractive alternatives. However, the non-trivial relationship between surplus gains and temperature profiles suggests that regulators should analyze carefully where to implement these structures.

6.5.3 On Carbon Emissions

A final element we explore in this analysis is how the different rate structures impact carbon emissions. A first observation is that not all technologies we consider are economical. Neither coal, nor biomass or nuclear are profitable. Perhaps one could have anticipated this outcome in light of the figures in table 6.1, which shows that geothermal, solar and wind dominate nuclear, biomass and coal. This is not a fair comparison, however, because these resources are of a different nature. While geothermal power plants have important geographic limitations, wind and solar are intermittent resources. Thus, one cannot discard a priori technologies with dominated economic characteristics.

A second observation is that some technologies do not change their total production profile or capacities across the rate scenarios. Consistently, those technologies do not alter their carbon emissions. This technologies include hydro and wind generating power plants. We expected hydro power generation to be invariant because it was exogenous in this analysis. The invariance of wind generation, on the other hand, is a outcome of the model.

Table 6.3 shows changes in capacity production and emissions with respect to the reference case. In addition, the table shows total change as a percentage of the Western Interconnection total for the FR scenario. We do not include unprofitable technologies nor technologies that do not vary across rate scenarios.

Agreeing with the basic insight of peak-load pricing, the total installed capacity decreases the most under real-time pricing, followed by TOU&CPP, TOU&DC and TOU. The order in terms of total installed capacity is not the same for total production. Indeed, the figures in Table 6.3 show that the order is somewhat reversed, with RTP increasing production the most. These changes, however, are a minor fraction of the total production of the Western Interconnection.

Table 6.3: Capacity, Production and Emissions Changes with Respect to FR Scenario

Metric	Tariff	Gas adv. CC	Gas adv. CT	Gas conv. CC	Solar	Total	Change relative to FR total
[MW-year]							%
Capacity	RTP	315	(8,561)	(177)	439	(7,984)	(6.09)
	TOU&CPP	(715)	(3,609)	(754)	2,449	(2,629)	(2.01)
	TOU&DC	(598)	(3,043)	(142)	1,463	(2,320)	(1.77)
	TOU	309	(1,600)	(642)	15	(1,918)	(1.46)
[GWh/year]							%
Production	RTP	2,197	(2,900)	156	1,620	1,074	0.17
	TOU&CPP	(6,932)	(610)	(1,177)	9,268	549	0.09
	TOU&DC	(4,356)	(793)	(33)	5,536	352	0.06
	TOU	1,693	(216)	(1,135)	55	398	0.06
[kt of CO ₂ eq/year]							%
Emissions	RTP	725	(1,508)	55	-	(728)	(0.62)
	TOU&CPP	(2,288)	(317)	(412)	-	(3,017)	(2.56)
	TOU&DC	(1,438)	(412)	(12)	-	(1,861)	(1.58)
	TOU	559	(112)	(397)	-	49	0.04

Even though production always changes positively, emissions do not. This is true for all rate scenarios with the exception of the TOU case. What happens is that solar production increases considerably in all but the TOU scenario. This suggests long-run complementarities between the demand responsiveness of the residential sector in California and solar generating plants in the Western Interconnection. Interestingly, RTP is not the rate that increases this complementarity the most. The combination between a TOU and a CPP program is the rate alternative that increases solar production and reduces emissions more notoriously.

6.6 Conclusions

We conduct an analysis of rate design in California's residential electricity sector. Beyond the applied insights, we contribute with a modeling framework to evaluate rate structures. The framework gives an important step towards bridging top down models of pricing and investment with bottom up models of household behavior. Building upon the theory of peak-load pricing, we illustrate how to modify the basic model to accommodate household heterogeneity, as well as the adoption of distributed energy resources and more general types of rate structures.

Our analysis seeks to quantify efficiency, distributional and environmental implications of rate structures that are plausible alternatives for California's future residential sector. The analysis shows that the average gains of implementing time-varying rates with respect to a simple flat rate program are rather mild, even in the real-time pricing scenario. Our results also show that factors

such as the presence of an air conditioning system and the exterior temperature profile can have a meaningful impact on the surplus gains that different rates generate on households. These two results combined suggest that defaulting all residential customers into a time-of-use rate structure, which is the current path California is following for the residential sector, may not be an ideal strategy. Targeting different rates to households with different appliance stocks and in different locations will likely be a superior policy.

References

- [1] MOSEK Modeling Manual, 2013.
- [2] Jan Paul Acton and Mitchell Bridger M. Welfare Analysis of Electricity Rate Changes, May 1983.
- [3] M. H. Albadi and E. F. El-Saadany. A summary of demand response in electricity markets. *Electric Power System Research*, 78(11):1989–1996, 2008.
- [4] Hunt Allcott. Rethinking real-time electricity pricing. *Resource and Energy Economics*, 33(4):820–842, November 2011.
- [5] L. Alvaro, S. Andy, and Z. Eugene, L. and Tongxin. Multistage adaptive robust optimization for the unit commitment problem. *Operations Research*, 2015.
- [6] Rafael J Avalos, Claudio A Canizares, and Miguel F Anjos. A practical voltage-stability-constrained optimal power flow. In *Proc. IEEE PES General Meeting*, pages 1–6. IEEE, 2008.
- [7] Marc Beaudin and Hamidreza Zareipour. Home energy management systems: A review of modelling and complexity. *Renewable and Sustainable Energy Reviews*, 45:318 – 335, 2015.
- [8] D. Bertsimas, E. Litvinov, S. Xu Andy, Z. Jinye, and Z. Tongxin. Adaptive robust optimization for the security constrained unit commitment problem. *IEEE Transactions on Power System*, 28(1), 2013.
- [9] M. Boiteux. Peak-load pricing. *The Journal of Business*, 33(2):157–179, April 1960.
- [10] James C Bonbright. *Principles of public utility rates*. Public Utilities Reports, Incorporated, 1988, 2nd edition, March 1988.
- [11] S. Borenstein. Effective and Equitable Adoption of Opt-In Residential Dynamic Electricity Pricing. *Review of Industrial Organization*, 42(2):127–160, 2013.
- [12] Severin Borenstein. The Long-Run Efficiency of Real-Time Electricity Pricing. *The Energy Journal*, 26(3):93–116, January 2005.
- [13] Severin Borenstein. Customer risk from real-time retail electricity pricing: Bill volatility and hedgability. Technical report, National Bureau of Economic Research, 2006.
- [14] Severin Borenstein. The Private Net Benefits of Residential Solar PV: The Role of Electricity Tariffs, Tax Incentives and Rebates. Working Paper 21342, National Bureau of Economic Research, July 2015.
- [15] Severin Borenstein and Stephen Holland. On the efficiency of competitive electricity markets with time-invariant retail prices. *The RAND Journal of Economics*, 36(3):pp. 469–493, 2005.

- [16] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- [17] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- [18] Stephen Boyd and Lieven Vandenbergh. Chapter 11: Interior-point methods. In *Convex optimization*. Cambridge University Press, illustrated, reprint edition, 2009.
- [19] Brattle, E3, BEAR, and Aspen. The impacts of a Regional ISO-Operated Power Market in California. Technical report, California Independent System Operator, July 2016.
- [20] Stephen J Brown and David Sumner Sibley. *The theory of public utility pricing*. Cambridge University Press, 1986.
- [21] Richard H. Byrd, Jorge Nocedal, and Richard A. Waltz. Knitro: An Integrated Package for Nonlinear Optimization. In G. Di Pillo and M. Roma, editors, *Large-Scale Nonlinear Optimization*, number 83 in Nonconvex Optimization and Its Applications, pages 35–59. Springer US, 2006. DOI: 10.1007/0-387-30065-1_4.
- [22] Claudio Cañizares, William Rosehart, Alberto Berizzi, and Cristian Bovo. Comparison of voltage security constrained optimal power flow techniques. In *Proc. IEEE Power Eng. Soc. Summer Meeting*, volume 3, pages 1680–1685. IEEE, 2001.
- [23] Florin Capitanescu. Assessing reactive power reserves with respect to operating constraints and voltage stability. *IEEE Trans. Power Syst.*, 26(4):2224–2234, 2011.
- [24] Florin Capitanescu, Thierry Van Cutsem, and Louis Wehenkel. Coupling optimization and dynamic simulation for preventive-corrective control of voltage instability. *IEEE Trans. Power Syst.*, 24(2):796–805, 2009.
- [25] Dennis W. Carlton. Peak load pricing with stochastic demand. *The American Economic Review*, 67(5):pp. 1006–1010, 1977.
- [26] M. Carrion, A. Conejo, and J. Arroyo. Forward contracting and selling price determination for a retailer. *IEEE Transactions on Power System*, 22(4):2105–2114, 2007.
- [27] F. A. Castro and D. S. Callaway. Optimal rate design in modern electricity sectors. Manuscript submitted for publication, 2016.
- [28] F. A. Castro, J.D. Lara, and D. S. Callaway. A mathematical programming approach to utility pricing. Manuscript in preparation, 2016.
- [29] Douglas W. Caves, Laurits R. Christensen, and Joseph A. Herriges. Consistency of residential customer response in time-of-use electricity pricing experiments. *Journal of Econometrics*, 26(1):179–203, September 1984.

- [30] Douglas W. Caves, Laurits R. Christensen, Philip E. Schoech, and Wallace Hendricks. A comparison of different methodologies in a case study of residential time-of-use electricity pricing: Costbenefit analysis. *Journal of Econometrics*, 26(12):17 – 34, 1984.
- [31] Hung-po Chao. Peak load pricing and capacity planning with demand and supply uncertainty. *The Bell Journal of Economics*, 14(1):pp. 179–190, 1983.
- [32] Hung-po Chao. Efficient pricing and investment in electricity markets with intermittent resources. *Energy Policy*, 39(7):3945 – 3953, 2011. Special Section: Renewable energy policy and development.
- [33] Miguel Chávez-Lugo, Claudio R Fuerte-Esquivel, Claudio A Cañizares, and Victor J Gutierrez-Martinez. Practical security boundary-constrained dc optimal power flow for electricity markets. *IEEE Trans. Power Syst.*, 31(5):3358–3368, 2016.
- [34] Z. Chen, L. Wu, and Y. Fu. Real-time demand response model. *IEEE Transactions on Smart Grid*, 3(4), 2012.
- [35] F. Clarke. *Optimization and Nonsmooth Analysis*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, January 1990. DOI: 10.1137/1.9781611971309 DOI: 10.1137/1.9781611971309.
- [36] Carleton Coffrin, Dan Gordon, and Paul Scott. Nesta, the nicta energy system test case archive. *arXiv preprint arXiv:1411.0359*, 2014.
- [37] A. J. Conejo, J. M. Morales, and L. Baringo. Real-time demand response model. *IEEE Transactions on Smart Grid*, 1(3), 2010.
- [38] Michael A. Crew, Chitru S. Fernando, and Paul R. Kleindorfer. The theory of peak-load pricing: A survey. *Journal of Regulatory Economics*, 8(3):215–248, November 1995.
- [39] Michael A. Crew and Paul R. Kleindorfer. Peak load pricing with a diverse technology. *The Bell Journal of Economics*, 7(1):pp. 207–231, 1976.
- [40] C. De Jonghe, B.F. Hobbs, and R. Belmans. Optimal generation mix with short-term demand response and wind penetration. *Power Systems, IEEE Transactions on*, 27(2):830–839, May 2012.
- [41] Stephan Dempe and Joydeep Dutta. Is bilevel programming a special case of a mathematical program with complementarity constraints? *Mathematical programming*, 131(1-2):37–48, 2012.
- [42] DOE. Demand reductions from the application of advance metering infrastructure, pricing programs and customer-based systems. Technical report, U.S. Department of Energy, December 2012.
- [43] DOE. Demand reductions from the application of advance metering infrastructure, pricing programs and customer-based systems. Technical report, US Department of Energy, December 2012.

- [44] DOE. Operations and maintenance savings from advance metering infrastructure. Technical report, U.S. Department of Energy, December 2012.
- [45] DOE. Operations and maintenance savings from advance metering infrastructure. Technical report, US Department of Energy, December 2012.
- [46] Jacques H. Drèze. Some postwar contributions of French economists to theory and public policy: With special emphasis on problems of resource allocation. *The American Economic Review*, 54(4):pp. 2–64, 1964.
- [47] EIA. Updated capital cost estimates for electricity generation plants. Technical report, US Energy Information Administration, April 2013.
- [48] EIA. Annual Energy Outlook 2016, with Projections to 2040. Technical report, US Energy Information Administration, August 2016.
- [49] Cherrelle Eid, Javier Reneses Guilln, Pablo Frasn, and Rudi Hakvoort. The economic effect of electricity net-metering with solar PV: Consequences for network cost recovery, cross subsidies and policy objectives. *Energy Policy*, 75:244–254, December 2014.
- [50] Tomaso Erseghe. Distributed optimal power flow using admm. *IEEE Transactions on Power Systems*, 29(5):2370–2380, 2014.
- [51] Yueyue Fan and Changzheng Liu. Solving Stochastic Transportation Network Protection Problems Using the Progressive Hedging-based Method. *Networks and Spatial Economics*, 10(2):193–208, June 2010.
- [52] Ahmad Faruqui, Dan Harris, and Ryan Hledik. Unlocking the 53 billion savings from smart meters in the eu: How increasing the adoption of dynamic tariffs could make or break the EUs smart grid investment. *Energy Policy*, 38(10):6222 – 6231, 2010. The socio-economic transition towards a hydrogen economy - findings from European research, with regular papers.
- [53] Ahmad Faruqui and Sanem Sergici. Household response to dynamic pricing of electricity: A survey of 15 experiments. *J Regul Econ*, 38(2):193–225, October 2010.
- [54] Ahmad Faruqui and Sanem Sergici. Dynamic pricing of electricity in the mid-Atlantic region: econometric results from the Baltimore gas and electric company experiment. *Journal of Regulatory Economics*, 40(1):82–109, 2011.
- [55] Ahmad Faruqui and Sanem Sergici. Arcturus: International evidence on dynamic pricing. *The Electricity Journal*, 26(7):55 – 65, 2013.
- [56] FERC. Assesment of demand response & advanced metering. Staff report Docket AD06-2-000, Federal Energy Regulatory Commission, August 2006.
- [57] R.S. Ferreira, L.A. Barroso, and M.M. Carvalho. Demand response models with correlated price data: A robust optimization approach. *Applied Energy*, 97, 2012.

- [58] Jos Fortuny-Amat and Bruce McCarl. A Representation and Economic Interpretation of a Two-Level Programming Problem. *The Journal of the Operational Research Society*, 32(9):783–792, 1981.
- [59] Manuel Frondel, Stephan Sommer, and Colin Vance. The burden of Germanys energy transition: An empirical analysis of distributional effects. *Economic Analysis and Policy*, 45:89–99, March 2015.
- [60] A. Ronald Gallant and Roger W. Koenker. Costs and benefits of peak-load pricing of electricity. *Journal of Econometrics*, 26(1):83–113, September 1984.
- [61] Giovanni Giallombardo and Daniel Ralph. Multiplier convergence in trust-region methods with application to convergence of decomposition methods for MPECs. *Mathematical Programming*, 112(2):335–369, April 2008.
- [62] Santiago Grijalva. Individual branch and path necessary conditions for saddle-node bifurcation voltage collapse. *IEEE Trans. Power Syst.*, 27(1):12–19, 2012.
- [63] Zhaomiao Guo and Yueyue Fan. A Stochastic Multi-agent Optimization Model for Energy Infrastructure Planning under Uncertainty in An Oligopolistic Market. *Networks and Spatial Economics*, pages 1–29, December 2016.
- [64] Roh. H and Lee. J. Residential demand response scheduling with multiclass appliances in the smart grid. *IEEE Transactions on Smart Grid*, 7(1), 2015.
- [65] T. G. Healy. Enernoc 2013 annual report. <http://investor.enernoc.com/annual-proxy.cfm>, 2013.
- [66] Karen Herter. Residential implementation of critical-peak pricing of electricity. *Energy Policy*, 35(4):2121–2130, April 2007.
- [67] Stephen P. Holland and Erin T. Mansur. Is Real-Time Pricing Green? The Environmental Impacts of Electricity Demand Variance. *The Review of Economics and Statistics*, 90(3):550–561, July 2008.
- [68] M. Hong, Z. Luo, and M. Razaviyayn. Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems. *SIAM Journal on Optimization*, 26(1):337–364, January 2016.
- [69] E. Philip Howrey and Hal R. Varian. Estimating the distributional impact of time-of-day pricing of electricity. *Journal of Econometrics*, 26(1):65–82, September 1984.
- [70] J.C.C.M. Huijben and G.P.J. Verbong. Breakthrough without subsidies? {PV} business model experiments in the netherlands. *Energy Policy*, 56:362 – 370, 2013.
- [71] IEA. Secure and efficient electricity supply during the transition to low carbon power system. Technical report, OECD, Paris, 2013.

- [72] IEA. Re-powering markets: Market design and regulation during the transition to low-carbon power systems. Technical report, OECD/IEA, 9 rue de la Federation, 75739 Paris Cedex 15, France, 2016.
- [73] Paul Joskow and Jean Tirole. Retail electricity competition. *The RAND Journal of Economics*, 37(4):799–815, 2006.
- [74] Paul Joskow and Jean Tirole. Reliability and competitive electricity markets. *The Rand Journal of Economics*, 38(1):60–84, 2007.
- [75] Paul L. Joskow. Contributions to the theory of marginal cost pricing. *The Bell Journal of Economics*, 7(1):pp. 197–206, 1976.
- [76] Paul L. Joskow. Regulation of natural monopoly. *Handbook of law and economics*, 2:1227–1348, 2007.
- [77] Paul L. Joskow and Catherine D. Wolfram. Dynamic pricing of electricity. *American Economic Review*, 102(3):381–85, 2012.
- [78] Alfred Edward Kahn. *The economics of regulation: Principles and institutions*. MIT press, 1988.
- [79] E. Karangelos and F. Bouffard. Towards full integration of demand-side resources in joint forward energy/reserve electricity markets. *IEEE Transactions on Power System*, 27(1), 2012.
- [80] P Kessel and H Glavitsch. Estimating the voltage stability of a power system. *IEEE Trans. Power Del.*, 1(3):346–354, 1986.
- [81] Balho H Kim and Ross Baldick. Coarse-grained distributed optimal power flow. *IEEE Transactions on Power Systems*, 12(2):932–939, 1997.
- [82] S. Kim and G.B. Giannakis. Scalable and robust demand response with mixed-integer constraints. *IEEE Transactions on Smart Grid*, 4(4), 2013.
- [83] T. T. Kim and H. V. Poor. Scheduling power consumption with price uncertainty. *IEEE Transactions on Smart Grid*, 2(2), 2011.
- [84] D.S. Kirschen. Demand-side view of electricity markets. *IEEE Transactions on Power Systems*, 18(2), 2003.
- [85] Burak Kocuk, Santanu S Dey, and X Andy Sun. Strong socp relaxations for the optimal power flow problem. *Operations Research*, 64(6):1177–1196, 2016.
- [86] Sameh KM Koudsi and Claudio A Canizares. Application of a stability-constrained optimal power flow to tuning of oscillation controls in competitive electricity markets. *IEEE Trans. Power Syst.*, 22(4):1944, 2007.
- [87] A. Grhan K  k, Kevin Shang, and afak Ycel. Impact of Electricity Pricing Policies on Renewable Energy Investments and Carbon Emissions. *Management Science*, December 2016.

- [88] VS Sravan Kumar, K Krishna Reddy, and D Thukaram. Coordination of reactive power in grid-connected wind farms for voltage stability enhancement. *IEEE Trans. Power Syst.*, 29(5):2381–2390, 2014.
- [89] N. Lappas and C. Gounaris. Robust optimization for decision-making under endogenous uncertainty. *Computers and Chemical Engineering*, 111, 2018.
- [90] J. Lazar and W. Gonzalez. Smart rate design for a smart future, 2015.
- [91] Lee A. Lillard and Dennis J. Aigner. Time-of-Day Electricity Consumption Response to Temperature and the Ownership of Air Conditioning Appliances. *Journal of Business & Economic Statistics*, 2(1):40–53, 1984.
- [92] Ovidiu Listes and Rommert Dekker. A Scenario AggregationBased Approach for Determining a Robust Airline Fleet Composition for Dynamic Capacity Allocation. *Transportation Science*, 39(3):367–382, August 2005.
- [93] Arne Løkketangen and David L. Woodruff. Progressive hedging and tabu search applied to mixed integer (0,1) multistage stochastic programming. *Journal of Heuristics*, 2(2):111–128, September 1996.
- [94] David G. Luenberger and Yinyu Ye. Chapter 10 Quasi-Newton methods. In *Linear and nonlinear programming*, volume 116. Springer, 2008.
- [95] Zhi-Quan Luo, Jong-Shi Pang, and Daniel Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, 1996.
- [96] Sindri Magnússon, Pradeep Chathuranga Weeraddana, and Carlo Fischione. A distributed approach for the optimal power-flow problem based on admm and sequential convex approximations. *IEEE Transactions on Control of Network Systems*, 2(3):238–253, 2015.
- [97] D. Marco and G. Ignacio. An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming*, 36, 1986.
- [98] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Number Book, Whole. Oxford University Press, New York, NY., 1995.
- [99] Johanna L. Mathieu, Mark E.H. Dyson, and Duncan S. Callaway. Resource and revenue potential of california residential load participation in ancillary services. *Energy Policy*, 80:76 – 87, 2015.
- [100] G. P. McCormick. Computability of global solutions to factorable nonconvex programs. *Mathematical Programming*, 10(1), 1976.
- [101] Federico Milano, Claudio A Cañizares, and Antonio J Conejo. Sensitivity-based security-constrained opf market clearing model. *IEEE Trans. Power Syst.*, 20(4):2051–2060, 2005.
- [102] Federico Milano, Claudio A Cañizares, and Marco Invernizzi. Multiobjective optimization for pricing system security in electricity markets. *IEEE Trans. Power Syst.*, 18(2):596–604, 2003.

- [103] Federico Milano, Claudio A Canizares, and Marco Invernizzi. Voltage stability constrained opf market models considering $n - 1$ contingency criteria. *Elec. Power Syst. Res.*, 74(1):27–36, 2005.
- [104] Frederic H. Murphy and Yves Smeers. Generation Capacity Expansion in Imperfectly Competitive Restructured Electricity Markets. *Operations research*, 53(4):646–661, 2005.
- [105] NARUC. Manual on distributed energy resources rate design compensation, 2016.
- [106] O. Nohadani and K. Sharma. Optimization under decision-dependent uncertainty. *SIAM Journal on Optimization*, 2018.
- [107] Richard P. O’Neill, Paul M. Sotkiewicz, Benjamin F. Hobbs, Michael H. Rothkopf, and William R. Stewart Jr. Efficient market-clearing prices in markets with nonconvexities. *European Journal of Operational Research*, 164(1):269–285, July 2005.
- [108] C. Palmgren, N. Stevens, M. Goldberg, R. Bames, and K. Rothkin. California Residential Appliance Saturation Survey. Technical Report CEC-200-2010-004, California Energy Commission, 2010.
- [109] John C. Panzar. A neoclassical approach to peak load pricing. *The Bell Journal of Economics*, 7(2):pp. 521–530, 1976.
- [110] Anthony Papavasiliou and Shmuel S. Oren. Multiarea Stochastic Unit Commitment for High Wind Penetration in a Transmission Constrained Network. *Operations research*, 61(3):578–592, 2013.
- [111] Richard W. Parks and David Weitzel. Measuring the consumer welfare effects of time-differentiated electricity prices. *Journal of Econometrics*, 26(12):35 – 64, 1984.
- [112] M. Parvania and M. Fotuhi-Firuzabad. Demand response scheduling by stochastic scuc. *IEEE Transactions on Smart Grid*, 1(1), 2010.
- [113] Andreas S Pedersen, Mogens Blanke, and Hjortur Jóhannsson. Convex relaxation of power dispatch for voltage stability improvement. In *Proc. IEEE Multi-Conference on Systems and Control*, pages 1528–1533. IEEE, 2015.
- [114] Qiuyu Peng and Steven H Low. Distributed optimal power flow algorithm for radial networks, i: Balanced single phase case. *IEEE Transactions on Smart Grid*, 9(1):111–121, 2018.
- [115] Charles F Phillips Jr. *The regulation of public utilities*. Public Utilities Reports, Incorporated, 1993, third edition, July 1993.
- [116] Alberto Del Pia, Santanu S. Dey, and Marco Molinaro. Mixed-integer quadratic programming is in NP. *Mathematical Programming*, 162(1-2):225–240, March 2017.
- [117] Iraj Rahimi Pordanjani, Yunfei Wang, and Wilsun Xu. Identification of critical components for voltage stability assessment using channel components transform. *IEEE Trans. Smart Grid*, 4(2):1122–1132, 2013.

- [118] J. E. Price and J. Goodin. Reduced network modeling of WECC as a market design prototype. In *Power and Energy Society General Meeting, 2011 IEEE*, pages 1–6, 2011. Conference Proceedings.
- [119] Henriquez. R, Wenzel. G, Olivares. D, and Negrete-Pincetic. M. Participation of demand response aggregators in electricity markets: Optimal portfolio management. *IEEE Transactions on Smart Grid*, 2017.
- [120] RAP. Electricity regulation in the us: A guide, 2011.
- [121] M Rastegar and M. Fotuhi Firuzabad. Outage management in residential demand response programs. *IEEE Transactions on Smart Grid*, 6(3), 2014.
- [122] Peter C. Reiss and Matthew W. White. Household Electricity Demand, Revisited. *The Review of Economic Studies*, 72(3):853–883, July 2005.
- [123] R. T. Rockafellar and Roger J.-B. Wets. Scenarios and Policy Aggregation in Optimization under Uncertainty. *Mathematics of Operations Research*, 16(1):119–147, 1991.
- [124] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.
- [125] Moon. S and Lee. J. Multi-residential demand response scheduling with multi-class appliances in smart grid. *IEEE Transactions on Smart Grid*, 2016.
- [126] Nan. S, Zhou. M, and Li. G. Optimal residential community demand response scheduling in smart grid. *Applied Energy*, 210, 2018.
- [127] Enzo E. Sauma and Shmuel S. Oren. Proactive planning and valuation of transmission investments in restructured electricity markets. *Journal of Regulatory Economics*, 30(3):261–290, November 2006.
- [128] Holger Scheel and Scholtes Stefan. Mathematical Programs with Complementarity Constraints: Stationarity, Optimality, and Sensitivity. *Mathematics of Operations Research*, 25(1):1–22, February 2000.
- [129] Winfried Schirotzek. *Nonsmooth analysis*. Springer Science & Business Media, 2007.
- [130] Ruggero Schleicher-Tappeser. How renewables will change electricity markets in the next five years. *Energy Policy*, 48:64–75, September 2012.
- [131] Lisa Schwartz, Max Wei, William Morrow, Jeff Deason, Steven Schiller, Greg Leventis, Sarah Smith, Woei Ling, Todd Levin, Steven Plotkin, Yan Zhou, and Joseph Teng. Electricity end uses, energy efficiency, and distributed energy resources baseline. study LBNL-1006983, Lawrence Berkeley National Laboratory, Berkeley, CA, US, January 2017.
- [132] John W Simpson-Porco, Florian Dörfler, and Francesco Bullo. Voltage collapse in complex power grids. *Nature Commun.*, 7:10790, 2016.

- [133] Ramteen Sioshansi. OR Forum Modeling the Impacts of Electricity Tariffs on Plug-In Hybrid Electric Vehicle Charging, Costs, and Emissions. *Operations Research*, 60(3):506–516, February 2012.
- [134] Tom Stanton. Distributed Energy Resources: Status Report on Evaluating Proposals and Practices for Electric Utility Rate Design. Status Report 15-08, National Regulatory Research Institute, 8611 Second Avenue, Suite 2C, Silver Spring, MD 20910, October 2015.
- [135] Peter O. Steiner. Peak loads and efficient pricing. *The Quarterly Journal of Economics*, 71(4):pp. 585–610, 1957.
- [136] Steven Stoft. *Power System Economics: Designing Markets for Electricity*. Number Book, Whole. IEEE Press, Piscataway, NJ., 2002.
- [137] Andy X Sun, Dzung T Phan, and Soumyadip Ghosh. Fully decentralized ac optimal power flow algorithms. In *Power and Energy Society General Meeting (PES), 2013 IEEE*, pages 1–5. IEEE, 2013.
- [138] Reza Takapoui, Nicholas Moehle, Stephen Boyd, and Alberto Bemporad. A simple effective heuristic for embedded mixed-integer quadratic programming. *International Journal of Control*, 0(0):1–11, April 2017.
- [139] Thomas N. Taylor, Peter M. Schwarz, and James E. Cochell. 24/7 Hourly Response to Electricity Real-Time Pricing with up to Eight Summers of Experience. *Journal of Regulatory Economics*, 27(3):235–262, May 2005.
- [140] Marco Todescato, John W Simpson-Porco, Florian Dörfler, Ruggero Carli, and Francesco Bullo. Online distributed voltage stress minimization by optimal feedback reactive power control. *IEEE Trans. Control Netw. Syst.*, 2017.
- [141] K. M. Tsui and S. C. Chan. Demand response optimization for smart home scheduling under real-time pricing. *IEEE Transactions on Smart Grid*, 3(4):1812–1821, 2012.
- [142] Krishna. V and Tsang. D. A two-stage approach for network constrained unit commitment problem with demand response. *IEEE Transactions on Smart Grid*, 9(2), 2018.
- [143] Thierry Van Cutsem and Costas Vournas. *Voltage stability of electric power systems*. Springer Science & Business Media, 2008.
- [144] Vaithianathan Venkatasubramanian, H Schattler, and John Zaborszky. Voltage dynamics: Study of a generator with voltage control, transmission, and matched mw load. *IEEE Trans. Autom. Control*, 37(11):1717–1733, 1992.
- [145] Aisma Vitina. Wind energy development in Denmark. In *IEA Wind Task 26 - Wind Technology, Cost, and Performance Trends in Denmark, Germany, Ireland, Norway, the European Union, and the United States: 2007-2012*, volume Chapter 1, pages 16–47. National Renewable Energy Laboratory, Golden, CO, USA, hand, m. m., ed. edition, June 2015.

- [146] Heinrich von Stackelberg. *The Theory of the Market Economy*. Oxford University Press, Oxford, England, (english translation of marktform und gleichgewicht, springer-verlag, berlin, 1934) edition, 1954.
- [147] Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Prog.*, 106(1):25–57, 2006.
- [148] Q. Wang, J. Wang, and Y. Guan. Stochastic unit commitment with uncertain demand response. *IEEE Transactions on Power Systems*, 28(1), 2013.
- [149] Yang Wang, Caisheng Wang, Feng Lin, Wenyuan Li, Le Yi Wang, and Junhui Zhao. Incorporating generator equivalent model into voltage stability analysis. *IEEE Trans. Power Syst.*, 28(4):4857–4866, 2013.
- [150] Zhaoyu Wang, Bai Cui, and Jianhui Wang. A necessary condition for power flow insolvability in power distribution systems with distributed generators. *IEEE Trans. Power Syst.*, 32(2):1440–1450, 2017.
- [151] Max Wei, James H. Nelson, Jeffery B. Greenblatt, Ana Mileva, Josiah Johnston, Michael Ting, Christopher Yang, Chris Jones, James E. McMahon, and Daniel M. Kammen. Deep carbon reductions in California require electrification and integration across economic sectors. *Environmental Research Letters*, 8(1):014038, 2013.
- [152] S. Wilcox and W. Marion. User manual for TMY3 data sets. Technical Report NREL/TP-581-43156, National Renewable Energy Laboratory, May 2008.
- [153] Robert D. Willig. Consumer’s Surplus Without Apology. *The American Economic Review*, 66(4):589–597, 1976.
- [154] Zhang. X, Shahidehpour. M, Alabdulwahab. A, and Abusorrah. A. Hourly electricity demand response in the stochastic day-ahead scheduling of coordinated electricity and natural gas networks. *IEEE Transactions on Power System*, 31(1), 2016.
- [155] Jane J. Ye. Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. *Journal of Mathematical Analysis and Applications*, 307(1):350 – 369, 2005.
- [156] Tarik Zabaoui, Louis-A Dessaint, and Innocent Kamwa. Preventive control approach for voltage stability improvement using voltage stability constrained optimal power flow based on static line voltage stability indices. *IET Gener. Transm. Distrib.*, 8(5):924–934, 2014.
- [157] C. Zhao, J. P. Watson J. Wang, and Y. Guan. Multi-stage robust unit commitment considering wind and demand response uncertainties. *IEEE Transactions on Power System*, 28(3), 2013.
- [158] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, Robert John Thomas, et al. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Trans. Power Syst.*, 26(1):12–19, 2011.

- [159] Gregor Zöttl. A framework of peak load pricing with strategic firms. *Operations Research*, 58(6):1637–1649, 2010.